CHAPTER 3

Probability, Statistics, and Information Theory

Randomness and uncertainty play an important role in science and engineering. Most spoken language processing problems can be characterized in a probabilistic framework. Probability theory and statistics provide the mathematical language to describe and analyze such systems.

The criteria and methods used to estimate the unknown probabilities and probability densities form the basis for estimation theory. Estimation theory forms the basics for parameter learning in pattern recognition. In this chapter, three widely used estimation methods are discussed. They are *minimum mean squared error estimation* (MMSE), *maximum likelihood estimation* (MLE), and *maximum posterior probability estimation* (MAP).

Significance testing is also important in statistics, which deals with the confidence of statistical inference, such as knowing whether the estimation of some parameter can be accepted with confidence. In pattern recognition, significance testing is extremely important for determining whether the observed difference between two different classifiers is real. In our coverage of significance testing, we describe various methods that are used in pattern recognition discussed in. Chapter 4.

Information theory was originally developed for efficient and reliable communication systems. It has evolved into a mathematical theory concerned with the very essence of the communication process. It provides a framework for the study of fundamental issues, such as the efficiency of information representation and the limitations in reliable transmission of information over a communication channel. Many of these problems are fundamental to spoken language processing.

3.1. Probability Theory

Probability theory deals with the averages of mass phenomena occurring sequentially or simultaneously. We often use probabilistic expressions in our day-to-day lives, such as when saying, *It is very likely that the Dow (Dow Jones Industrial index) will hit 12,000 points next month*, or, *The chance of scattered showers in Seattle this weekend is high.* Each of these expressions is based upon the concept of the probability, or the likelihood, which some specific event will occur.

Probability can be used to represent the degree of confidence in the outcome of some actions (observations), which are not definite. In probability theory, the term *sample space, S,* is used to refer to the collection (set) of all possible outcomes. An *event* refers to a subset of the sample space or a collection of outcomes. The *probability of event A* denoted as $P(A)$, can be interpreted as the *relative frequency* with which the event *A* would occur if the process were repeated a large number of times under similar conditions. Based on this interpretation, $P(A)$ can be computed simply by counting the total number, N_S, of all observations and the number of observations N_A whose outcome belongs to the event *A*. That is,

$$P(A) = \frac{N_A}{N_S} \tag{3.1}$$

$P(A)$ is bounded between zero and one, i.e.,

$$0 \le P(A) \le 1 \text{ for all } A \tag{3.2}$$

The lower bound of probability $P(A)$ is zero when the event set *A* is an empty set. On the other hand, the upper bound of probability $P(A)$ is one when the event set *A* happens to be *S*.

If there are *n* events $A_1, A_2, \cdots A_n$ in *S* such that $A_1, A_2, \cdots A_n$ are disjoint and $\bigcup_{i=1}^{n} A_i = S$, events $A_1, A_2, \cdots A_n$ are said to form a *partition* of *S*. The following obvious equation forms a fundamental axiom for probability theory.

$$P(A_1 \cup A_2 \cup \ldots A_n) = \sum_{i=1}^{n} P(A_i) = 1 \tag{3.3}$$

Based on the definition in Eq. (3.1), the *joint probability* of event *A* and event *B* occurring concurrently is denoted as $P(AB)$ and can be calculated as:

$$P(AB) = \frac{N_{AB}}{N_S} \tag{3.4}$$

3.1.1. Conditional Probability And Bayes' Rule

It is useful to study the way in which the probability of an event *A* changes after it has been learned that some other event *B* has occurred. This new probability denoted as $P(A \mid B)$ is called the *conditional probability* of event *A* given that event *B* has occurred. Since the set of those outcomes in *B* that also result in the occurrence of *A* is exactly the set *AB* as illustrated in Figure 3.1, it is natural to define the conditional probability as the proportion of the total probability $P(B)$ that is represented by the joint probability $P(AB)$. This leads to the following definition:

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{N_{AB}/N_S}{N_B/N_S} \tag{3.5}$$

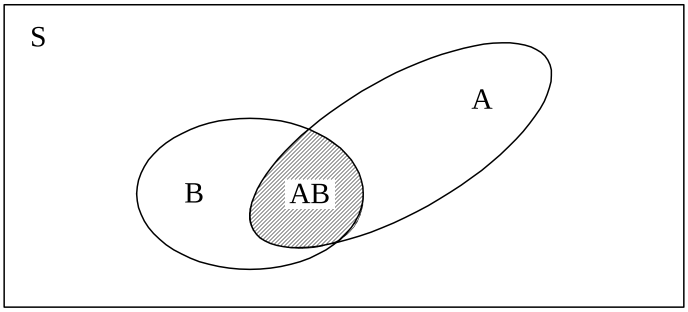

Figure 3.1 The intersection AB represents where the joint event *A* and *B* occurs concurrently.

Based on the definition of conditional probability, the following expressions can be easily derived.

$$P(AB) = P(A \mid B)P(B) = P(B \mid A)P(A) \tag{3.6}$$

Equation (3.6) is the simple version of the *chain rule*. The chain rule, which can specify a joint probability in terms of multiplication of several cascaded conditional probabilities, is often used to decompose a complicated joint probabilistic problem into a sequence of stepwise conditional probabilistic problems. Eq. (3.6) can be converted to such a general chain:

$$P(A_1A_2 \cdots A_n) = P(A_n \mid A_1 \cdots A_{n-1}) \cdots P(A_2 \mid A_1)P(A_1) \tag{3.7}$$

When two events, *A* and *B*, are independent of each other, in the sense that the occurrence or of either of them has no relation to and no influence on the occurrence of the other, it is obvious that the conditional probability $P(B \mid A)$ equals to the unconditional probability $P(B)$. It follows that the joint probability $P(AB)$ is simply the product of $P(A)$ and $P(B)$ if *A* and *B*, are independent.

If the *n* events $A_1, A_2, \cdots A_n$ form a partition of *S* and *B* is any event in *S* as illustrated in Figure 3.2, the events $A_1B, A_2B, \cdots A_nB$ form a partition of *B*. Thus, we can rewrite:

$$B = A_1B \cup A_2B \cup \cdots \cup A_nB \tag{3.8}$$

Since $A_1B, A_2B, \cdots A_nB$ are disjoint,

$$P(B) = \sum_{k=1}^{n} P(A_kB) \tag{3.9}$$

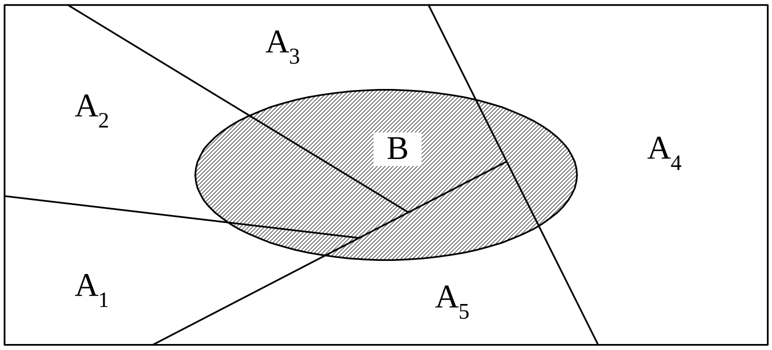

Figure 3.2 The intersections of *B* with partition events $A_1, A_2, \cdots A_n$.

Equation (3.9) is called the *marginal probability* of event *B*, where the probability of event *B* is computed from the sum of joint probabilities.

According to the chain rule, Eq. (3.6), $P(A_iB) = P(A_i)P(B \mid A_i)$, it follows that

$$P(B) = \sum_{k=1}^{n} P(A_k)P(B \mid A_k) \tag{3.10}$$

Combining Eqs. (3.5) and (3.10), we get the well-known *Bayes' rule*:

$$P(A_i \mid B) = \frac{P(A_iB)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{k=1}^{n} P(B \mid A_k)P(A_k)} \tag{3.11}$$

Bayes' rule is the basis for pattern recognition that is described in Chapter 4.

3.1.2. Random Variables

Elements in a sample space may be numbered and referred to by the numbers given. A variable *X* that specifies the numerical quantity in a sample space is called a *random variable*. Therefore, a random variable *X* is a function that maps each possible outcome *s* in the sample space *S* onto real numbers $X(s)$. Since each event is a subset of the sample space, an event is represented as a set of $\{s\}$ which satisfies $\{s \mid X(s) = x\}$. We use capital letters to denote random variables and lower-case letters to denote fixed values of the random variable. Thus, the probability that $X = x$ is denoted as:

$$P(X = x) = P(s \mid X(s) = x) \tag{3.12}$$

A random variable *X* is a *discrete* random variable, or *X* has a *discrete distribution*, if *X* can take only a finite number *n* of different values $x_1, x_2, \cdots, x_n$, or at most, an infinite sequence of different values $x_1, x_2, \cdots$. If the random variable *X* is a discrete random variable, the *probability function* (p.f.) or *probability mass function* (p.m.f.) of *X* is defined to be the function *p* such that for any real number *x*,

$$p_X(x) = P(X = x) \tag{3.13}$$

For the cases in which there is no confusion, we drop the subscription *X* for $p_X(x)$. The sum of probability mass over all values of the random variable is equal to unity.

$$\sum_{k=1}^{n} p(x_i) = \sum_{k=1}^{n} P(X = x_i) = 1 \tag{3.14}$$

The marginal probability, chain rule and Bayes' rule can also be rewritten with respect to random variables.

$$p_X(x_i) = P(X = x_i) = \sum_{k=1}^{m} P(X = x_i, Y = y_k) = \sum_{k=1}^{m} P(X = x_i \mid Y = y_k) P(Y = y_k) \tag{3.15}$$

$$\begin{aligned} &P(X_1 = x_1, \cdots, X_n = x_n) = \\ &\quad P(X_n = x_n \mid X_1 = x_1, \cdots, X_{n-1} = x_{n-1}) \cdots P(X_2 = x_2 \mid X_1 = x_1) P(X_1 = x_1) \end{aligned} \tag{3.16}$$

$$P(X = x_i \mid Y = y) = \frac{P(X = x_i, Y = y)}{P(Y = y)} = \frac{P(Y = y \mid X = x_i) P(X = x_i)}{\sum_{k=1}^{n} P(Y = y \mid X = x_k) P(X = x_k)} \tag{3.17}$$

In a similar manner, if the random variables *X* and *Y* are statistically independent, they can be represented as:

$$P(X = x_i, Y = y_j) = P(X = x_i) P(Y = y_j) = p_X(x_i) p_Y(y_j) \ \forall \text{ all } i \text{ and } j \tag{3.18}$$

A random variable X is a *continuous* random variable, or X has a *continuous distribution,* if there exists a nonnegative function f, defined on the real line, such that for an interval A,

$$P(X \in A) = \int_A f_X(x)dx \tag{3.19}$$

The function f_X is called the *probability density function* (abbreviated p.d.f.) of X. We drop the subscript X for f_X if there is no ambiguity. As illustrated in Figure 3.3, the area of shaded region is equal to the value of $P(a \le X \le b)$

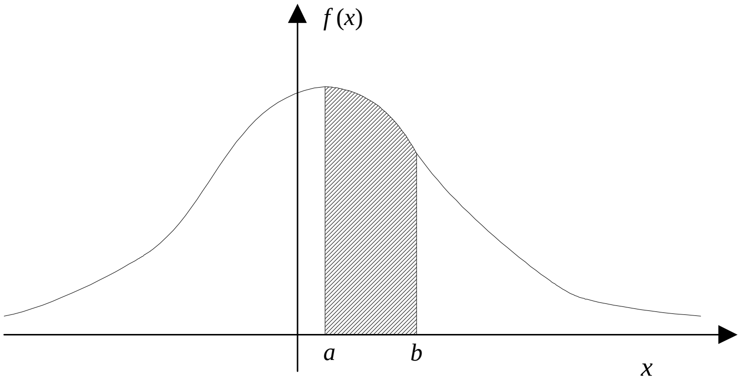

Figure 3.3 An example of p.d.f. The area of the shaded region is equal to the value of $P(a \le X \le b)$.

Every p.d.f must satisfy the following two requirements.

$$\begin{gathered} f(x) \ge 0 \text{ for } -\infty \le x \le \infty \text{ and} \\ \int_{-\infty}^{\infty} f(x)dx = 1 \end{gathered} \tag{3.20}$$

The marginal probability, chain rule, and Bayes' rule can also be rewritten with respect to continuous random variables:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)dy \tag{3.21}$$

$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = f_{X_n|X_1,\cdots,X_{n-1}}(x_n \mid x_1,\cdots,x_{n-1}) \cdots f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1) \tag{3.22}$$

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y \mid x) f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y \mid x) f_X(x)dx} \tag{3.23}$$

The *distribution function* or *cumulative distribution function F* of a discrete or continuous random variable *X* is a function defined for all real number *x* as follows:

$$F(x) = P(X \le x) \quad \text{for } -\infty \le x \le \infty \tag{3.24}$$

For continuous random variables, It follows that:

$$F(x) = \int_{-\infty}^{x} f_X(x)dx \tag{3.25}$$

$$f_X(x) = \frac{dF(x)}{dx} \tag{3.26}$$

3.1.3. Mean and Variance

Suppose that a discrete random variable *X* has a p.f. *f*(*x*); the *expectation* or *mean* of *X* is defined as follows:

$$E(X) = \sum_x xf(x) \tag{3.27}$$

Similarly, if a continuous random variable *X* has a p.d.f. *f*, the *expectation* or *mean* of *X* is defined as follows:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \tag{3.28}$$

In physics, the mean is regarded as the center of mass of the probability distribution. The expectation can also be defined for any function of the random variable *X*. If *X* is a continuous random variable with p.d.f. *f*, then the expectation of any function $g(X)$ can be defined as follows:

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)dx \tag{3.29}$$

The expectation of a random variable is a linear operator. That is, it satisfies both additivity and homogeneity properties:

$$E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b \tag{3.30}$$

where $a_1, \cdots, a_n, b$ are constants

Equation (3.30) is valid regardless of whether or not the random variables $X_1, \cdots, X_n$ are independent.

Suppose that *X* is a random variable with mean $\mu = E(X)$. The *variance* of *X* denoted as $Var(X)$ is defined as follows:

$$Var(X) = \sigma^2 = E\left[(X-\mu)^2\right] \tag{3.31}$$

where σ, the nonnegative square root of the variance is known as the *standard deviation* of random variable *X*. Therefore, the variance is also often denoted as σ^2.

The variance of a distribution provides a measure of the spread or dispersion of the distribution around its mean μ. A small value of the variance indicates that the probability distribution is tightly concentrated around μ, and a large value of the variance typically indicates the probability distribution has a wide spread around μ. Figure 3.4 illustrates three different Gaussian distributions[1] with the same mean, but different variances.

The variance of random variable *X* can be computed in the following way:

$$Var(X) = E(X^2) - \left[E(X)\right]^2 \tag{3.32}$$

In physics, the expectation $E(X^k)$ is called the k^{th} moment of *X* for any random variable *X* and any positive integer *k*. Therefore, the variance is simply the difference between the second moment and the square of the first moment.

The variance satisfies the following additivity property, if random variables *X* and *Y* are independent:

$$Var(X+Y) = Var(X) + Var(Y) \tag{3.33}$$

However, it does not satisfy the homogeneity property. Instead for constant *a*,

$$Var(aX) = a^2 Var(X) \tag{3.34}$$

Since it is clear that $Var(b) = 0$ for any constant *b*, we have an equation similar to Eq. (3.30) if random variables $X_1, \cdots, X_n$ are independent.

$$Var(a_1X_1 + \cdots + a_nX_n + b) = a_1^2 Var(X_1) + \cdots + a_n^2 Var(X_n) \tag{3.35}$$

Conditional expectation can also be defined in a similar way. Suppose that *X* and *Y* are discrete random variables and let $f(y \mid x)$ denote the conditional p.f. of *Y* given $X = x$, then the conditional expectation $E(Y \mid X)$ is defined to be the function of *X* whose value $E(Y \mid x)$ when $X = x$ is

$$E_{Y|X}(Y \mid X = x) = \sum_y y f_{Y|X}(y \mid x) \tag{3.36}$$

For continuous random variables *X* and *Y* with $f_{Y|X}(y \mid x)$ as the conditional p.d.f. of *Y* given $X = x$, the conditional expectation $E(Y \mid X)$ is defined to be the function of *X* whose value $E(Y \mid x)$ when $X = x$ is

[1] We describe Gaussian distributions in Section 3.1.7

$$E_{Y|X}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy \tag{3.37}$$

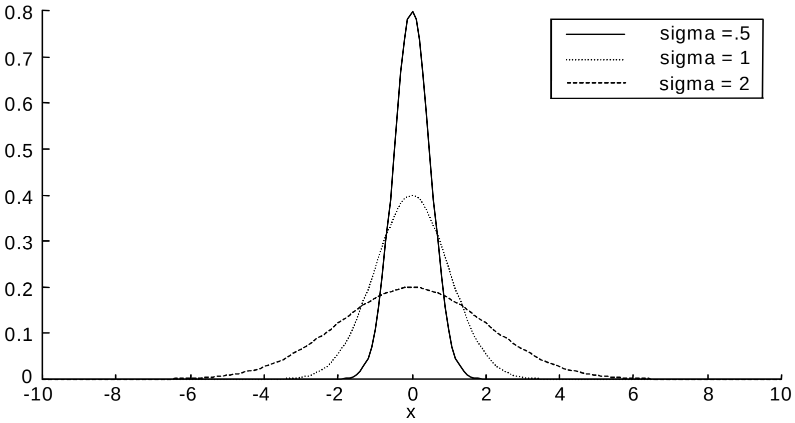

Figure 3.4 Three Gaussian distributions with same mean μ, but different variances, 0.5, 1.0 ,and 2.0, respectively. The distribution with a large value of the variance has a wide spread around the mean μ.

Since $E(Y \mid X)$ is a function of random variable *X*, it itself is a random variable whose probability distribution can be derived from the distribution of *X*. It can be shown that

$$E_X\left[E_{Y|X}(Y \mid X)\right] = E_{X,Y}(Y) \tag{3.38}$$

More generally, suppose that *X* and *Y* have a continuous joint distribution and that $g(x, y)$ is any arbitrary function of *X* and *Y*. The conditional expectation $E\left[g(X,Y) \mid X\right]$ is defined to be the function of *X* whose value $E\left[g(X,Y) \mid x\right]$ when $X = x$ is

$$E_{Y|X}\left[g(X,Y) \mid X = x\right] = \int_{-\infty}^{\infty} g(x, y) f_{Y|X}(y \mid x) dy \tag{3.39}$$

Equation (3.38) can also be generalized into the following equation:

$$E_X\left\{E_{Y|X}\left[g(X,Y) \mid X\right]\right\} = E_{X,Y}\left[g(X,Y)\right] \tag{3.40}$$

Finally, it is worthwhile to introduce *median* and *mode*. A median of the distribution of *X* is defined to be a point *m*, such that $P(X \le m) \ge 1/2$ and $P(X \ge m) \ge 1/2$. Thus, the median *m* divides the total probability into two equal parts, i.e., the probability to the left of *m* and the probability to the right of *m* are exactly $1/2$.

Suppose a random variable X has either a discrete distribution with p.f. $p(x)$ or continuous p.d.f. $f(x)$; a point ϖ is called the mode of the distribution if $p(x)$ or $f(x)$ attains the maximum value at the point ϖ. A distribution can have more than one modes.

3.1.3.1. The Law of Large Numbers

The concept of *sample mean* and *sample variance* is important in statistics because most statistical experiments involve sampling. Suppose that the random variables $X_1, \cdots, X_n$ form a random sample of size n from some distribution for which the mean is μ and the variance is σ^2. In other words, the random variables $X_1, \cdots, X_n$ are *independent identically distributed* (often abbreviated by i.i.d.) and each has mean μ and variance σ^2. Now if we denote $\bar{X}_n$ as the arithmetic average of the n observations in the sample, then

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) \tag{3.41}$$

$\bar{X}_n$ is a random variable and is referred to as *sample mean*. The mean and variance of $\bar{X}_n$ can be easily derived based on the definition.

$$E(\bar{X}_n) = \mu \quad \text{and} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \tag{3.42}$$

Equation (3.42) states that the mean of sample mean is equal to mean of the distribution, while the variance of sample mean is only $1/n$ times the variance of the distribution. In other words, the distribution of $\bar{X}_n$ will be more concentrated around the mean μ than was the original distribution. Thus, the sample mean is closer to μ than is the value of just a single observation X_i from the given distribution.

The *law of large numbers* is one of most important theorems in probability theory. Formally, it states that the sample mean $\bar{X}_n$ converges to the mean μ in probability, that is,

$$\lim_{n\to\infty} P\left(|\bar{X}_n - \mu| < \varepsilon\right) = 1 \text{ for any given number } \varepsilon > 0 \tag{3.43}$$

The law of large numbers basically implies that the sample mean is an excellent estimate of the unknown mean of the distribution when the sample size n is large.

3.1.4. Covariance and Correlation

Let X and Y be random variables having a specific joint distribution, and $E(X) = \mu_X$, $E(Y) = \mu_Y$, $Var(X) = \sigma_X^2$, and $Var(Y) = \sigma_Y^2$. The *covariance* of *X* and *Y*, denoted as $Cov(X,Y)$, is defined as follows:

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = Cov(Y,X) \tag{3.44}$$

In addition, the *correlation coefficient* of *X* and *Y*, denoted as ρ_{XY}, is defined as follows:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.45}$$

It can be shown that $\rho(X,Y)$ should be bound within $\left[-1\ldots1\right]$, that is,

$$-1 \le \rho(X,Y) \le 1 \tag{3.46}$$

X and *Y* are said to be *positively correlated* if $\rho_{XY} > 0$, *negatively correlated* if $\rho_{XY} < 0$, and *uncorrelated* if $\rho_{XY} = 0$. It can also be shown that $Cov(X,Y)$ and ρ_{XY} must have the same sign; that is, both are positive, negative, or zero at the same time. When $E(XY) = 0$, the two random variables are called *orthogonal*.

There are several theorems pertaining to the basic properties of covariance and correlation. We list here the most important ones:

Theorem 1 For any random variables *X* and *Y*

$$Cov(X,Y) = E(XY) - E(X)E(Y) \tag{3.47}$$

Theorem 2 If *X* and *Y* are independent random variables, then

$$Cov(X,Y) = \rho_{XY} = 0$$

Theorem 3 Suppose *X* is a random variable and *Y* is a linear function of *X* in the form of $Y = aX + b$ for some constant *a* and *b*, where $a \ne 0$. If $a > 0$, then $\rho_{XY} = 1$. If $a < 0$, then $\rho_{XY} = -1$. Sometimes, ρ_{XY} is referred to as the amount of linear dependency between random variables *X* and *Y*.

Theorem 4 For any random variables *X* and *Y*,

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \tag{3.48}$$

Theorem 5 If $X_1, \cdots, X_n$ are random variables, then

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} Cov(X_i, X_j) \tag{3.49}$$

3.1.5. Random Vectors and Multivariate Distributions

When a random variable is a vector rather than a scalar, it is called a random vector and we often use boldface variable like $\mathbf{X} = (X_1, \cdots, X_n)$ to indicate that it is a random vector. It is said that *n* random variables $X_1, \cdots, X_n$ have a *discrete joint distribution* if the random vector $\mathbf{X} = (X_1, \cdots, X_n)$ can have only a finite number or an infinite sequence of different values $(x_1, \cdots, x_n)$ in R^n. The joint p.f. of $X_1, \cdots, X_n$ is defined to be the function $f_\mathbf{X}$ such that for any point $(x_1, \cdots, x_n) \in R^n$,

$$f_\mathbf{X}(x_1, \cdots, x_n) = P(X_1 = x_1, \cdots, X_n = x_n) \tag{3.50}$$

Similarly, it is said that *n* random variables $X_1, \cdots, X_n$ have a *continuous joint distribution* if there is a nonnegative function *f* defined on R^n such that for any subset $A \subset R^n$,

$$P\left[(X_1, \cdots, X_n) \in A\right] = \int_A \cdots \int f_\mathbf{X}(x_1, \cdots, x_n) dx_1 \cdots dx_n \tag{3.51}$$

The *joint distribution function* can also be defined similarly for *n* random variables $X_1, \cdots, X_n$ as follows:

$$F_\mathbf{X}(x_1, \cdots, x_n) = P(X_1 \le x_1, \cdots, X_n \le x_n) \tag{3.52}$$

The concept of mean and variance for a random vector can be generalized *into mean vector* and *covariance matrix*. Supposed that **X** is an *n*-dimensional random vector with components $X_1, \cdots, X_n$, under matrix representation, we have

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \tag{3.53}$$

The *expectation (mean) vector* $E(\mathbf{X})$ of random vector **X** is an *n*-dimensional vector whose components are the expectations of the individual components of **X**, that is,

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix} \tag{3.54}$$

The *covariance matrix* $Cov(\mathbf{X})$ of random vector **X** is defined to be an $n \times n$ matrix such that the element in the i^{th} row and j^{th} column is $Cov(X_i, Y_j)$, that is,

$$Cov(\mathbf{X}) = \begin{bmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_n) \\ \vdots & & \vdots \\ Cov(X_n, X_1) & \cdots & Cov(X_n, X_n) \end{bmatrix} = E\left[\left[X - E(X)\right]\left[X - E(X)\right]^t\right] \quad (3.55)$$

It should be emphasized that the *n* diagonal elements of the covariance matrix $Cov(\mathbf{X})$ are actually the variances of $X_1, \cdots, X_n$. Furthermore, since covariance is symmetric, i.e., $Cov(X_i, X_j) = Cov(X_j, X_i)$, the covariance matrix $Cov(\mathbf{X})$ must be a symmetric matrix.

There is an important theorem regarding the mean vector and covariance matrix for a linear transformation of the random vector **X**. Suppose **X** is an *n*-dimensional vector as specified by Eq. (3.53), with mean vector $E(\mathbf{X})$ and covariance matrix $Cov(\mathbf{X})$. Now, assume **Y** is a *m*-dimensional random vector which is a linear transform of random vector **X** by the relation: $\mathbf{Y} = \mathbf{AX} + \mathbf{B}$, where **A** is a $m \times n$ transformation matrix whose elements are constants, and **B** is a *m*-dimensional constant vector. Then we have the following two equations:

$$E(\mathbf{Y}) = \mathbf{A}E(\mathbf{X}) + \mathbf{B} \quad (3.56)$$

$$Cov(\mathbf{Y}) = \mathbf{A}Cov(\mathbf{X})\mathbf{A}^t \quad (3.57)$$

3.1.6. Some Useful Distributions

In the following two sections, we will introduce several useful distributions that are widely used in applications of probability and statistics, particularly in spoken language systems.

3.1.6.1. Uniform Distributions

The simplest distribution is uniform distribution where the p.f. or p.d.f. is a constant function. For uniform discrete random variable *X*, which only takes possible values from $\{x_i \mid 1 \le i \le n\}$, the p.f. for *X* is

$$P(X = x_i) = \frac{1}{n} \qquad 1 \le i \le n \quad (3.58)$$

For uniform continuous random variable *X*, which only takes possible values from real interval $[a, b]$, the p.d.f. for *X* is

$$f(x) = \frac{1}{b-a} \qquad a \le x \le b \tag{3.59}$$

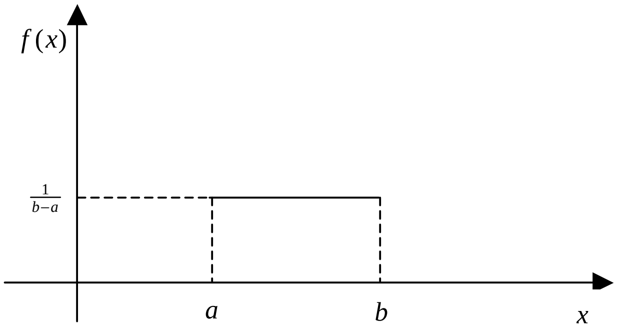

Figure 3.5 A uniform distribution for p.d.f. in Eq. (3.59)

3.1.6.2. Binomial Distributions

The *binomial distribution* is used to describe binary-decision events. For example, suppose that a single coin toss will produce the head with probability p and produce the tail with probability $1-p$. Now, if we toss the same coin n times and let X denote the number of heads observed, then the random variable X has the following *binomial* p.f.:

$$P(X = x) = f(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{3.60}$$

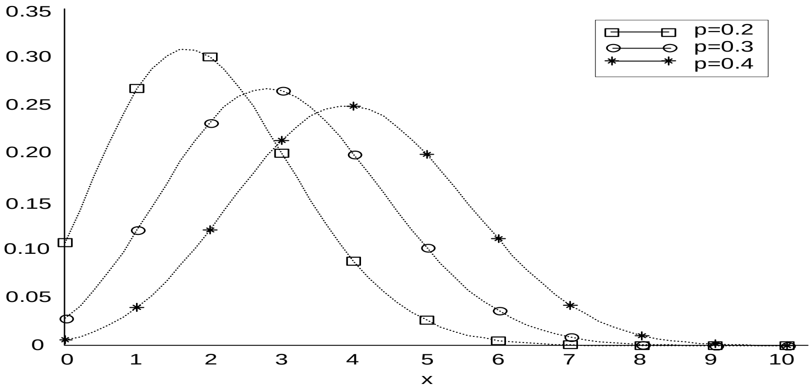

Figure 3.6 Three binomial distributions with p=0.2, 0.3 and 0.4.

It can be shown that the mean and variance of a binomial distribution are:

$$E(X) = np \tag{3.61}$$

$$Var(X) = np(1-p) \tag{3.62}$$

Figure 3.6 illustrates three binomial distributions with p = 0.2, 0.3 and 0.4.

3.1.6.3. Geometric Distributions

The geometric distribution is related to the binomial distribution. As in the independent coin toss example, the head-up has a probability p and the tail-up has a probability $1-p$. The geometric distribution is to model the time until a tail-up appears. Let the random variable X be the time (the number of tosses) until the first tail-up is shown. The p.d.f. of X is in the following form:

$$P(X = x) = f(x \mid p) = p^{x-1}(1-p) \quad x = 1, 2, \ldots \text{ and } 0 < p < 1 \tag{3.63}$$

The mean and variance of a geometric distribution are given by:

$$E(X) = \frac{1}{1-p} \tag{3.64}$$

$$Var(X) = \frac{1}{(1-p)^2} \tag{3.65}$$

One example for the geometric distribution is the distribution of the state duration for a hidden Markov model, as described in Chapter 8. Figure 3.7 illustrates three geometric distributions with p = 0.2, 0.3 and 0.4.

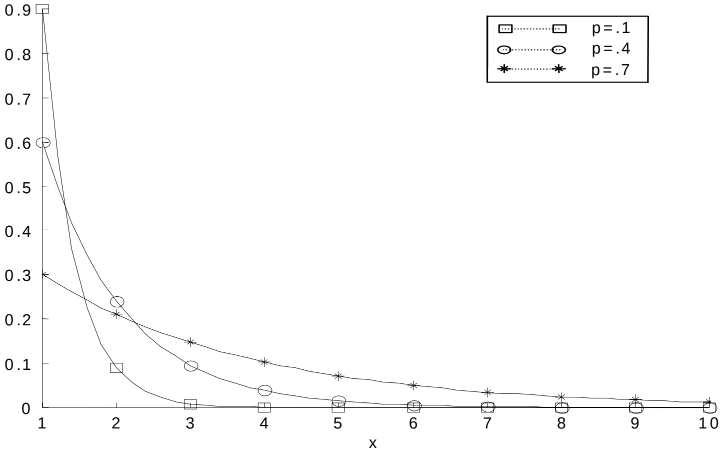

Figure 3.7 Three geometric distributions with different parameter p.

3.1.6.4. Multinomial Distributions

Suppose that a bag contains balls of *k* different colors, where the proportion of the balls of color *i* is p_i. Thus, $p_i > 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^{k} p_i = 1$. Now suppose that *n* balls are randomly selected from the bag and there are enough balls ($> n$) of each color. Let X_i denote the number of selected balls that are of color *i*. The random vector $\mathbf{X} = (X_1, \ldots, X_k)$ is said to have a *multinomial distribution* with parameters *n* and $\mathbf{p} = (p_1, \ldots p_k)$. For a vector $\mathbf{x} = (x_1, \ldots x_k)$, the p.f. of **X** has the following form:

$$P(\mathbf{X} = \mathbf{x}) = f(\mathbf{x} \mid n, \mathbf{p}) = \begin{cases} \dfrac{n!}{x_1!, \ldots x_k!} p_1^{x_1}, \ldots p_k^{x_k} & \text{where } x_{\mathrm{i}} \geq 0 \ \forall i = 1, \ldots, k \\ & \text{and } x_1 + \cdots + x_k = n \\ 0 & \text{otherwise} \end{cases} \tag{3.66}$$

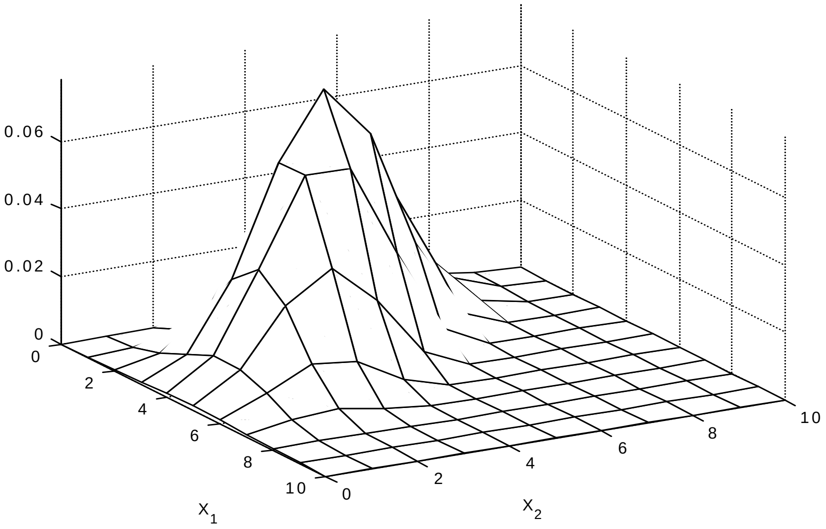

Figure 3.8 A multinomial distribution with *n*=10, $p_1 = 0.2$ and $p_2 = 0.3$

It can be shown that the mean, variance and covariance of the multinomial distribution are:

$$E(X_i) = np_i \text{ and } Var(X_i) = np_i(1 - p_i) \ \forall i = 1, \ldots, k \tag{3.67}$$

$$Cov(X_i, X_j) = -np_i p_j \tag{3.68}$$

Figure 3.8 shows a multinomial distribution with n = 10, $p_1 = 0.2$ and $p_2 = 0.3$. Since there are only two free parameters x_1 and x_2, the graph is illustrated only using x_1 and x_2 as axis. Multinomial distributions are typically used with the χ^2 test that is one of the most widely used goodness-of-fit hypotheses testing procedures described in Section 3.3.3.

3.1.6.5. Poisson Distributions

Another popular discrete distribution is *Poisson distribution*. The random variable *X* has a Poisson distribution with mean λ $(\lambda > 0)$ if the p.f. of *X* has the following form:

$$P(X = x) = f(x \mid \lambda) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} & \text{for } x\text{=0,1,2,}\ldots \\ 0 & \text{otherwise} \end{cases} \tag{3.69}$$

The mean and variance of a Poisson distribution are the same and equal λ:

$$E(X) = Var(X) = \lambda \tag{3.70}$$

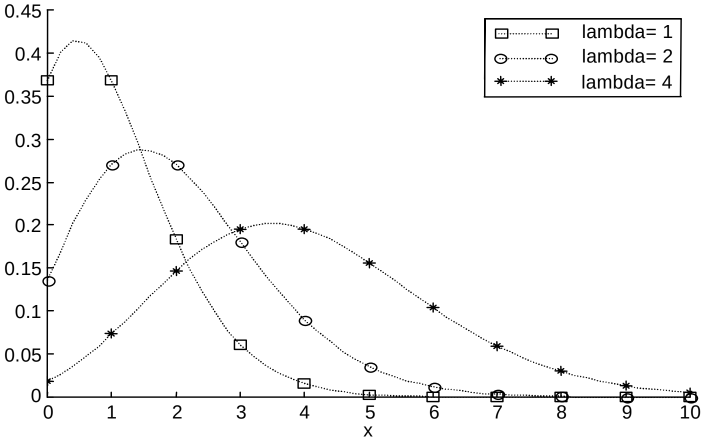

Figure 3.9 Three Poisson distributions with λ = 1, 2, and 4.

Figure 3.9 illustrates three Poisson distributions with $\lambda = 1$, 2, and 4. The Poisson distribution is typically used in queuing theory, where *x* is the total number of occurrences of some phenomenon during a fixed period of time or within a fixed region of space. Examples include the number of telephone calls received at a switchboard during a fixed period of time. In speech recognition, the Poisson distribution is used to model the duration for a phoneme.

3.1.6.6. Gamma Distributions

A continuous random variable *X* is said to have a *gamma distribution* with parameters α and β ($\alpha > 0$ and $\beta > 0$) if *X* has a continuous p.d.f. of the following form:

$$f(x \mid \alpha, \beta) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x>0 \\ 0 & x \le 0 \end{cases} \quad (3.71)$$

where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad (3.72)$$

It can be shown that the function Γ is a factorial function when α is a positive integer.

$$\Gamma(n) = \begin{cases} (n-1)! & n = 2,3,\ldots \\ 1 & n=1 \end{cases} \quad (3.73)$$

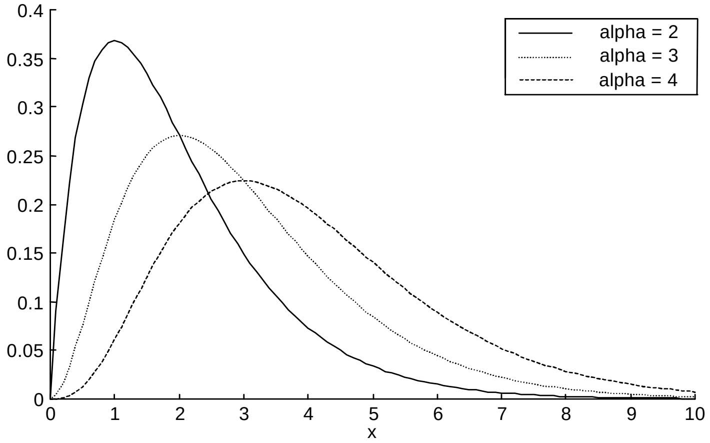

Figure 3.10 Three Gamma distributions with $\beta = 1.0$ and $\alpha = 2.0$, 3.0, and 4.0.

The mean and variance of a gamma distribution are:

$$E(X) = \frac{\alpha}{\beta} \text{ and } Var(X) = \frac{\alpha}{\beta^2} \quad (3.74)$$

Figure 3.10 illustrates three gamma distributions with $\beta = 1.0$ and $\alpha = 2.0$, 3.0, and 4.0. There is an interesting theorem associated with gamma distributions. If the random variables $X_1, \ldots, X_k$ are independent and each random variable X_i has a gamma distribution with parameters α_i and β, then the sum $X_1 + \cdots + X_k$ also has a gamma distribution with parameters $\alpha_1 + \cdots + \alpha_k$ and β.

A special case of gamma distribution is called *exponential distribution*. A continuous random variable *X* is said to have an *exponential distribution* with parameters β ($\beta > 0$) if *X* has a continuous p.d.f. of the following form:

$$f(x \mid \beta) = \begin{cases} \beta e^{-\beta x} & x > 0 \\ 0 & x \le 0 \end{cases} \tag{3.75}$$

It is clear that the exponential distribution is a gamma distribution with $\alpha = 1$. The mean and variance of the exponential distribution are:

$$E(X) = \frac{1}{\beta} \text{ and } Var(X) = \frac{1}{\beta^2} \tag{3.76}$$

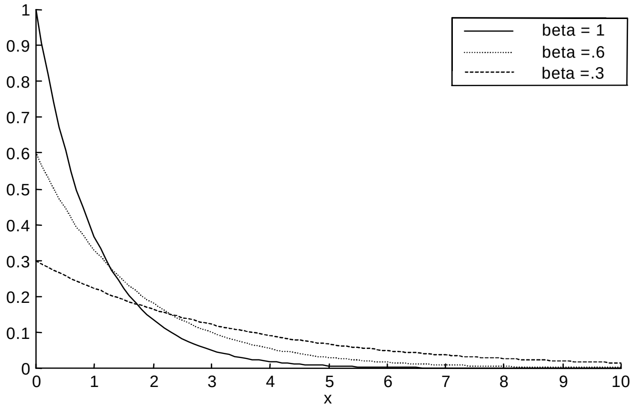

Figure 3.11 Three exponential distributions with β = 1.0, 0.6 and 0.3.

Figure 3.11 shows three exponential distributions with $\beta = 1.0$, 0.6, and 0.3. The exponential distribution is often used in queuing theory for the distributions of the duration of a service or the inter-arrival time of customers. It is also used to approximate the distribution of the life of a mechanical component.

3.1.7. Gaussian Distributions

Gaussian distribution is by far the most important probability distribution mainly because many scientists have observed that the random variables studied in various physical experiments (including speech signals), often have distributions that are approximately Gaussian. The Gaussian distribution is also referred to as normal distribution. A continuous random variable *X* is said to have a *Gaussian distribution* with mean μ and variance σ^2 ($\sigma > 0$) if *X* has a continuous p.d.f. in the following form:

$$f(x \mid \mu, \sigma^2) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \tag{3.77}$$

It can be shown that μ and σ^2 are indeed the mean and the variance for the Gaussian distribution. Some examples of Gaussian can be found in Figure 3.4.

The use of Gaussian distributions is justified by the *Central Limit Theorem*, which states that observable events considered to be a consequence of many unrelated causes with no single cause predominating over the others, tend to follow the Gaussian distribution [6].

It can be shown from Eq. (3.77) that the Gaussian $f(x \mid \mu, \sigma^2)$ is symmetric with respect to $x = \mu$. Therefore, μ is both the mean and the median of the distribution. Moreover, μ is also the mode of the distribution, i.e., the p.d.f. $f(x \mid \mu, \sigma^2)$ attains its maximum at the mean point $x = \mu$.

Several Gaussian p.d.f.'s with the same mean μ, but different variances are illustrated in Figure 3.4. Readers can see that the curve has a *bell* shape. The Gaussian p.d.f. with a small variance has a high peak and is very concentrated around the mean μ, whereas the Gaussian p.d.f., with a large variance, is relatively flat and is spread out more widely over the *x*-axis.

If the random variable *X* is a *Gaussian distribution* with mean μ and variance σ^2, then any linear function of *X* also has a Gaussian distribution. That is, if $Y = aX + b$, where *a* and *b* are constants and $a \neq 0$, *Y* has a Gaussian distribution with mean $a\mu + b$ and variance $a^2\sigma^2$. Similarly, the sum $X_1 + \cdots + X_k$ of independent random variables $X_1, \ldots, X_k$, where each random variable X_i has a Gaussian distribution, is also a Gaussian distribution.

3.1.7.1. Standard Gaussian Distributions

The Gaussian distribution with mean 0 and variance 1, denoted as $N(0,1)$, is called the *standard Gaussian distribution* or *unit Gaussian distribution*. Since the linear transformation of a Gaussian distribution is still a Gaussian distribution, the behavior of a Gaussian distribution can be solely described using a standard Gaussian distribution. If the random variable

X is a Gaussian distribution with mean μ and variance σ^2, that is, $X \sim N(\mu, \sigma^2)$, it can be shown that

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1) \tag{3.78}$$

Based on Eq. (3.78), the following property can be shown:

$$P(|X - \mu| \le k\sigma) = P(|Z| \le k) \tag{3.79}$$

Equation (3.79) demonstrates that every Gaussian distribution contains the same total amount of probability within any fixed number of standard deviations of its mean.

3.1.7.2. The Central Limit Theorem

If random variables $X_1, \ldots, X_n$ are i.i.d. according to a common distribution function with mean μ and variance σ^2, then as the random sample size *n* approaches ∞, the following random variable has a distribution converging to the standard Gaussian distribution:

$$Y_n = \frac{n(\overline{X}_n - \mu)}{\sqrt{n\sigma^2}} \sim N(0,1) \tag{3.80}$$

where $\overline{X}_n$ is the sample mean of random variables $X_1, \ldots, X_n$ as defined in Eq. (3.41).

Based on Eq. (3.80), the sample mean random variable $\overline{X}_n$ can be approximated by a Gaussian distribution with mean μ and variance σ^2 / n.

The central limit theorem above is applied to i.i.d. random variables $X_1, \ldots, X_n$. A. Liapounov in 1901 derived another central limit theorem for independent but not necessarily identically distributed random variables $X_1, \ldots, X_n$. Suppose $X_1, \ldots, X_n$ are independent random variables and $E(|X_i - \mu_i|^3) < \infty$ for $1 \le i \le n$; the following random variable will converge to standard Gaussian distribution when $n \to \infty$.

$$Y_n = (\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i) / \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{1/2} \tag{3.81}$$

In other words, the sum of random variables $X_1, \ldots, X_n$ can be approximated by a Gaussian distribution with mean $\sum_{i=1}^{n} \mu_i$ and variance $\left(\sum_{i=1}^{n} \sigma_i^2 \right)^{1/2}$.

Both central limit theorems essentially state that regardless of their original individual distributions, the sum of many independent random variables (effects) tends to be distributed like a Gaussian distribution as the number of random variables (effects) becomes large.

3.1.7.3. Multivariate Mixture Gaussian Distributions

When $\mathbf{X} = (X_1, \ldots, X_n)$ is an n-dimensional continuous random vector, the multivariate Gaussian p.d.f. has the following form:

$$f(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \tag{3.82}$$

where $\boldsymbol{\mu}$ is the n-dimensional mean vector, $\boldsymbol{\Sigma}$ is the $n \times n$ covariance matrix, and $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{\mu} = E(\mathbf{x}) \tag{3.83}$$

$$\boldsymbol{\Sigma} = E\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t \right] \tag{3.84}$$

More specifically, the *i-j*th element σ_{ij}^2 of covariance matrix $\boldsymbol{\Sigma}$ can be specified as follows:

$$\sigma_{ij}^2 = E\left[(x_i - \mu_i)(x_j - \mu_j) \right] \tag{3.85}$$

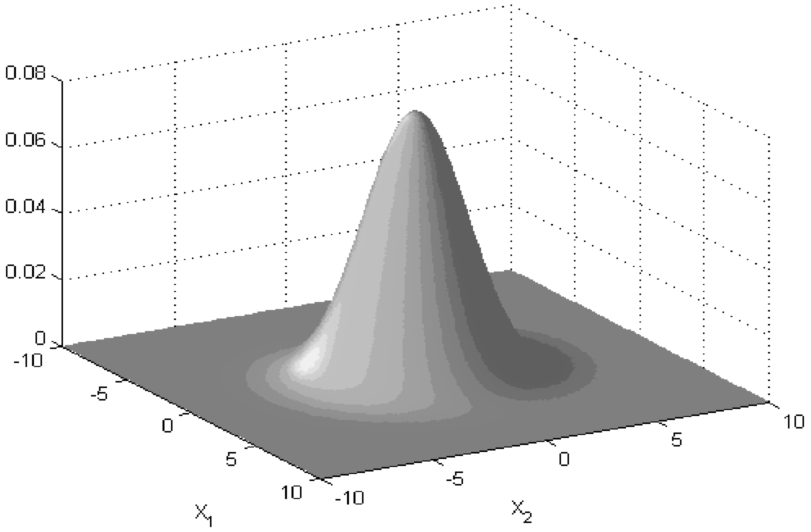

Figure 3.12 A two-dimensional multivariate Gaussian distribution with independent random variables x_1 and x_2 that have the same variance.

If $X_1, \ldots, X_n$ are independent random variables, the covariance matrix $\mathbf{\Sigma}$ is reduced to diagonal covariance where all the off-diagonal entries are zero. The distribution can be regarded as *n* independent scalar Gaussian distributions. The joint p.d.f. is the product of all the individual scalar Gaussian p.d.f.. Figure 3.12 shows a two-dimensional multivariate Gaussian distribution with independent random variables x_1 and x_2 with the same variance. Figure 3.13 shows another two-dimensional multivariate Gaussian distribution with independent random variables x_1 and x_2 that have different variances.

Although Gaussian distributions are unimodal,[2] more complex distributions with multiple local maxima can be approximated by Gaussian mixtures:

$$f(\mathbf{x}) = \sum_{k=1}^{K} c_k N_k(\mathbf{x}; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \tag{3.86}$$

where c_k, the mixture weight associated with *k*th Gaussian component are subject to the following constraint:

$$c_k \geq 0 \ \text{ and } \ \sum_{k=1}^{K} c_k = 1$$

Gaussian mixtures with enough mixture components can approximate any distribution. Throughout this book, most continuous probability density functions are modeled with Gaussian mixtures.

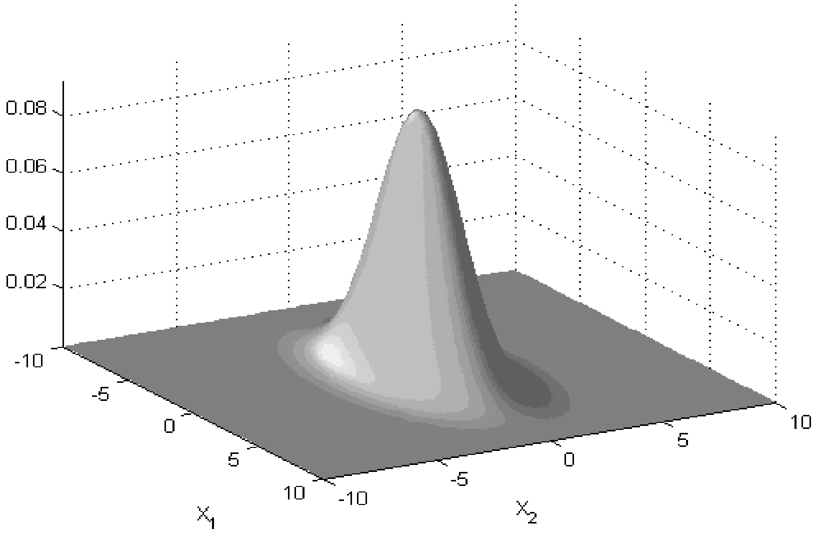

Figure 3.13 Another two-dimensional multivariate Gaussian distribution with independent

[2] A unimodal distribution has a single maximum (bump) for the distribution. For Gaussian distribution, the maximum occurs at the mean.

random variable x_1 and x_2 which have different variances.

3.1.7.4. χ^2 Distributions

A gamma distribution with parameters α and β is defined in Eq. (3.71). For any given positive integer *n*, the gamma distribution for which $\alpha = n/2$ and $\beta = 1/2$ is called the χ^2 *distribution* with *n* degrees of freedom. It follows from Eq. (3.71) that the p.d.f. for the χ^2 distribution is

$$f(x\,|\,n) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} & x>0 \\ 0 & x \le 0 \end{cases} \tag{3.87}$$

χ^2 distributions are important in statistics because they are closely related to random samples of Gaussian distribution. They are widely applied in many important problems of statistical inference and hypothesis testing. Specifically, if the random variables $X_1, \ldots, X_n$ are independent and identically distributed, and if each of these variables has a standard Gaussian distribution, then the sum of square $X_1^2 + \ldots + X_n^2$ can be proved to have a χ^2 distribution with *n* degree of freedom. Figure 3.14 illustrates three χ^2 distributions with $n = 2$, 3 and 4.

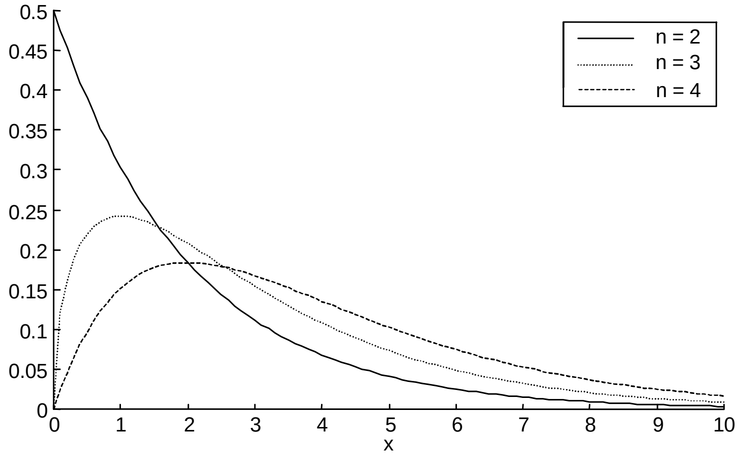

Figure 3.14 Three χ^2 distributions with $n = 2$, 3, and 4.

The mean and variance for the χ^2 distribution are

$$E(X) = n \text{ and } Var(X) = 2n \tag{3.88}$$

Following the additivity property of the gamma distribution, the χ^2 distribution also has the additivity property. That is, if the random variables $X_1, \ldots, X_n$ are independent and if X_i has a χ^2 distribution with k_i degrees of freedom, the sum $X_1 + \ldots + X_n$ has a χ^2 distribution with $k_1 + \ldots + k_n$ degrees of freedom.

3.1.7.5. Log-Normal Distribution

Let *x*, be a Gaussian random variable with mean μ_x and standard deviation σ_x, then

$$y = e^x \tag{3.89}$$

follows a *log-normal* distribution

$$f(y \mid \mu_x, \sigma_x) = \frac{1}{y\sigma_x\sqrt{2\pi}} \exp\left\{-\frac{(\ln y - \mu_x)^2}{2\sigma_x^2}\right\} \tag{3.90}$$

shown in Figure 3.15, and whose mean is given by

$$\begin{aligned} \mu_y &= E\{y\} = E\{e^x\} = \int_{-\infty}^{\infty} \exp\{x\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\} dx \\ &= \int_{-\infty}^{\infty} \exp\left\{\mu_x + \sigma_x^2/2\right\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-(\mu_x+\sigma_x^2)^2}{2\sigma_x^2}\right\} dx = \exp\left\{\mu_x + \sigma_x^2/2\right\} \end{aligned} \tag{3.91}$$

where we have rearranged the quadratic form of *x* and made use of the fact that the total probability mass of a Gaussian is 1. Similarly, the second order moment of *y* is given by

$$\begin{aligned} E\{y^2\} &= \int_{-\infty}^{\infty} \exp\{2x\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\} dx \\ &= \int_{-\infty}^{\infty} \exp\left\{2\mu_x + 2\sigma_x^2\right\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-(\mu_x+2\sigma_x^2)^2}{2\sigma_x^2}\right\} dx = \exp\left\{2\mu_x + 2\sigma_x^2\right\} \end{aligned} \tag{3.92}$$

and thus the variance of *y* is given by

$$\sigma_y^2 = E\{y^2\} - \left(E\{y\}\right)^2 = \mu_y^2\left(\exp\left\{\sigma_x^2\right\} - 1\right) \tag{3.93}$$

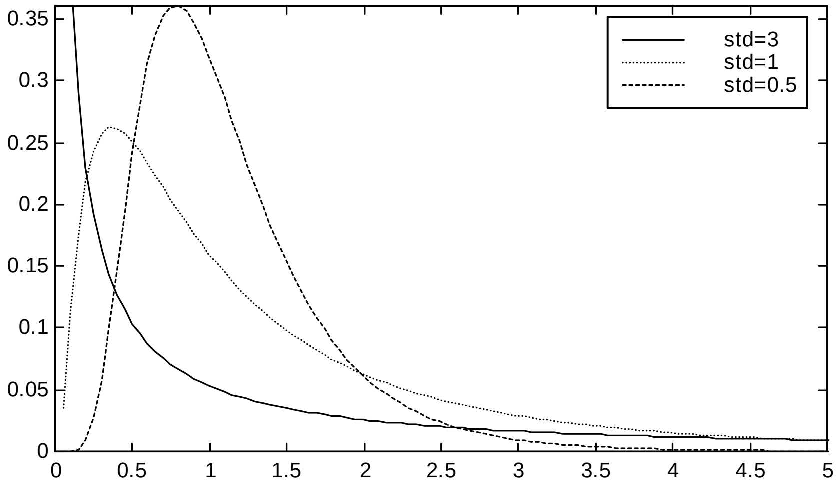

Figure 3.15 Lognormal distribution for $\mu_x = 0$ and $\sigma_x = 3$, 1 and 0.5 according to Eq. (3.90).

Similarly, if **x** is a Gaussian random vector with mean $\boldsymbol{\mu}_{\mathbf{x}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$, then random vector $\mathbf{y} = e^{\mathbf{x}}$ is log-normal with mean and covariance matrix [8] given by

$$\begin{aligned}&\boldsymbol{\mu}_{\mathrm{y}}[i] = \exp\left\{\boldsymbol{\mu}_{\mathbf{x}}[i] + \boldsymbol{\Sigma}_{\mathbf{x}}[i,i]/2\right\}\\&\boldsymbol{\Sigma}_{\mathrm{y}}[i,j] = \boldsymbol{\mu}_{\mathrm{y}}[i]\boldsymbol{\mu}_{\mathrm{y}}[j]\left(\exp\left\{\boldsymbol{\Sigma}_{\mathrm{x}}[i,j]\right\} - 1\right)\end{aligned} \tag{3.94}$$

using a similar derivation as in Eqs. (3.91) to (3.93).

3.2. Estimation Theory

Estimation theory and *significance testing* are two most important theories and methods of *statistical inference*. In this section, we describe estimation theory while significance testing is covered in the next section. A problem of statistical inference is one in which data generated in accordance with some unknown probability distribution must be analyzed, and some type of inference about the unknown distribution must be made. In a problem of statistical inference, any characteristic of the distribution generating the experimental data, such as the mean μ and variance σ^2 of a Gaussian distribution, is called a parameter of the distribution. The set Ω of all possible values of a *parameter* Φ or a group of parameters $\Phi_1, \Phi_2, \ldots, \Phi_n$ is called the *parameter space*. In this section we focus on how to estimate the parameter Φ from sample data.

Before we describe various estimation methods, we introduce the concept and nature of the estimation problems. Suppose that a set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ is i.i.d. according to a p.d.f. $p(x|\Phi)$ where the value of the parameter Φ is unknown. Now, suppose also that the value of Φ must be estimated from the observed values in the sample. An *estimator* of the parameter Φ, based on the random variables $X_1, X_2, \ldots, X_n$, is a real-valued function $\theta(X_1, X_2, \ldots, X_n)$ that specifies the estimated value of Φ for each possible set of values of $X_1, X_2, \ldots, X_n$. That is, if the sample values of $X_1, X_2, \ldots, X_n$ turn out to be $x_1, x_2, \ldots, x_n$, then the estimated value of Φ will be $\theta(x_1, x_2, \ldots, x_n)$.

We need to distinguish between *estimator*, *estimate*, and *estimation*. An estimator $\theta(X_1, X_2, \ldots, X_n)$ is a function of the random variables, whose probability distribution can be derived from the joint distribution of $X_1, X_2, \ldots, X_n$. On the other hand, an estimate is a specific value $\theta(x_1, x_2, \ldots, x_n)$ of the estimator that is determined by using some specific sample values $x_1, x_2, \ldots, x_n$. Estimation is usually used to indicate the process of obtaining such an estimator for the set of random variables or an estimate for the set of specific sample values. If we use the notation $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ to represent the vector of random variables and $\mathbf{x} = \left\{x_1, x_2, \ldots, x_n\right\}$ to represent the vector of sample values, an estimator can be denoted as $\theta(\mathbf{X})$ and an estimate $\theta(\mathbf{x})$. Sometimes we abbreviate an estimator $\theta(\mathbf{X})$ by just the symbol θ.

In the following four sections we describe and compare three different estimators (estimation methods). They are *minimum mean square estimator*, *maximum likelihood estimator*, and *Bayes estimator*. The first one is often used to estimate the random variable itself, while the latter two are used to estimate the parameters of the distribution of the random variables.

3.2.1. Minimum/Least Mean Squared Error Estimation

Minimum mean squared error (MMSE) estimation and *least squared error* (LSE) estimation are important methods for random variable since the goal (minimize the squared error) is an intuitive one. In general, two random variables *X* and *Y* are i.i.d. according to some p.d.f. $f_{X,Y}(x, y)$. Suppose that we perform a series of experiments and observe the value of *X*. We want to find a transformation $\hat{Y} = g(X)$ such that we can predict the value of the random variable *Y*. The following quantity can measure the goodness of such a transformation.

$$E(Y - \hat{Y})^2 = E(Y - g(X))^2 \tag{3.95}$$

This quantity is called *mean squared error* (MSE) because it is the mean of the squared error of the predictor $g(X)$. The criterion of minimizing the mean squared error is a good one for picking the predictor $g(X)$. Of course, we usually specify the class of function *G*, from which $g(X)$ may be selected. In general, there is a parameter vector $\mathbf{\Phi}$ associated with the function $g(X)$, so the function can be expressed as $g(X, \mathbf{\Phi})$. The process to find the parameter vector $\hat{\mathbf{\Phi}}_{MMSE}$ that minimizes the mean of the squared error is called *minimum mean squared error estimation* and $\hat{\mathbf{\Phi}}_{MMSE}$ is called the *minimum mean squared error estimator*. That is,

$$\hat{\mathbf{\Phi}}_{MMSE} = \arg\min_{\Phi} \left[E\left[(Y - g(X, \mathbf{\Phi}))^2 \right] \right] \tag{3.96}$$

Sometimes, the joint distribution of random variables *X* and *Y* is not known. Instead, samples of (*x*,*y*) pairs may be observable. In this case, the following criterion can be used instead,

$$\mathbf{\Phi}_{LSE} = \arg\min_{\Phi} \sum_{i=1}^{n} \left[y_i - g(x_i, \mathbf{\Phi}) \right]^2 \tag{3.97}$$

The argument of the minimization in Eq. (3.97) is called *sum-of-squared-error* (SSE) and the process of finding the parameter vector $\hat{\mathbf{\Phi}}_{LSE}$, which satisfies the criterion is called *least squared error estimation* or *minimum squared error estimation*. LSE is a powerful mechanism for curve fitting, where the function $g(x, \mathbf{\Phi})$ describes the observation pairs (x_i, y_i) . In general, there are more points (*n*) than the number of free parameters in function $g(x, \Phi)$, so the fitting is over-determined. Therefore, no exact solution exists, and LSE fitting becomes necessary.

It should be emphasized that MMSE and LSE are actually very similar and share similar properties. The quantity in Eq. (3.97) is actually *n* times the sample mean of the squared error. Based on the law of large numbers, when the joint probability $f_{X,Y}(x, y)$ is uniform or the number of samples approaches to infinity, MMSE and LSE are equivalent.

For the class of functions, we consider the following three cases:

- Constant functions, i.e.,

$$G_c = \left\{ g(x) = c, c \in R \right\} \tag{3.98}$$

- Linear functions, i.e.,

$$G_l = \left\{ g(x) = ax + b, \quad a, b \in R \right\} \tag{3.99}$$

- Other non-linear functions G_{nl}

3.2.1.1. MMSE/LSE for Constant Functions

When $\hat{Y} = g(x) = c$, Eq. (3.95) becomes

$$E(Y - \hat{Y})^2 = E(Y - c)^2 \tag{3.100}$$

To find the MMSE estimate for *c*, we take the derivatives of both sides in Eq. (3.100) with respect to *c* and equate it to 0. The MMSE estimate c_{MMSE} is given as

$$c_{MMSE} = E(Y) \tag{3.101}$$

and the minimum mean squared error is exactly the variance of *Y*, $Var(Y)$.

For the LSE estimate of *c*, the quantity in Eq. (3.97) becomes

$$\min \sum_{i=1}^{n} \left[y_i - c \right]^2 \tag{3.102}$$

Similarly, the LSE estimate c_{LSE} can be obtained as follows:

$$c_{LSE} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{3.103}$$

The quantity in Eq. (3.103) is the sample mean.

3.2.1.2. MMSE and LSE For Linear Functions

When $\hat{Y} = g(x) = ax + b$, Eq. (3.95) becomes

$$e(a,b) = E(Y - \hat{Y})^2 = E(Y - ax - b)^2 \tag{3.104}$$

To find the MMSE estimate of *a* and *b*, we can first set

$$\frac{\partial e}{\partial a} = 0, \text{ and } \frac{\partial e}{\partial b} = 0 \tag{3.105}$$

and solve the two linear equations. Thus, we can obtain

$$a = \frac{\operatorname{cov}(X,Y)}{Var(X)} = \rho_{XY} \frac{\sigma_Y}{\sigma_X} \tag{3.106}$$

$$b = E(Y) - \rho_{XY} \frac{\sigma_Y}{\sigma_X} E(X) \tag{3.107}$$

For LSE estimation, we assume that the sample **x** is a *d*-dimensional vector for generality. Assuming we have *n* sample-vectors $(\mathbf{x}_i, y_i) = (x_i^1, x_i^2, \cdots, x_i^d, y_i)$, $i = 1 \ldots n$, a linear function can be represented as

$$\hat{\mathbf{Y}} = \mathbf{XA} \text{ or } \begin{pmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \vdots \\ \mathrm{y}_\mathrm{n} \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^d \\ 1 & x_2^1 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \cdots & x_n^d \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} \tag{3.108}$$

The sum of squared error can then be represented as

$$e(\mathbf{A}) = \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2 = \sum_{i=1}^{n} \left(\mathbf{A}^t \mathbf{x}_i - y_i \right)^2 \tag{3.109}$$

A closed-form solution of the LSE estimate of **A** can be obtained by taking the gradient of $e(\mathbf{A})$,

$$\nabla e(\mathbf{A}) = \sum_{i=1}^{n} 2(\mathbf{A}^t \mathbf{x}_i - y_i)\mathbf{x}_i = 2\mathbf{X}^t(\mathbf{XA} - \mathbf{Y}) \tag{3.110}$$

and equating it to zero. This yields the following equation:

$$\mathbf{X}^t \mathbf{XA} = \mathbf{X}^t \mathbf{Y} \tag{3.111}$$

Thus the LSE estimate $\mathbf{A}_{LSE}$ will be of the following form:

$$\mathbf{A}_{LSE} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y} \tag{3.112}$$

$(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$ in Eq. (3.112) is also refereed to as the *pseudo-inverse* of **X** and is sometimes denoted as $\mathbf{X}^\perp$.

When $\mathbf{X}^t \mathbf{X}$ is singular or some boundary conditions cause the LSE estimation in Eq. (3.112) to be unattainable, some numeric methods can be used to find an approximate solution. Instead of minimizing the quantity in Eq. (3.109), one can minimize the following quantity:

$$e(\mathbf{A}) = \| \mathbf{XA} - \mathbf{Y} \|^2 + \alpha \| \mathbf{X} \|^2 \tag{3.113}$$

Following a similar procedure, one can obtain the LSE estimate to minimize the quantity above in the following form.

$$\mathbf{A}^*_{LSE} = (\mathbf{X}^t \mathbf{X} + \alpha \mathbf{I})^{-1} \mathbf{X}^t \mathbf{Y} \tag{3.114}$$

The LSE solution in Eq. (3.112) can be used for polynomial functions too. In the problem of polynomial curve fitting using the least square criterion, we are aiming to find the coefficients $\mathbf{A} = (a_0, a_1, a_2, \cdots, a_d)^t$ that minimize the following quantity:

$$\min_{a_0, a_1, a_2, \cdots, a_d} E(Y - \hat{Y})^2 \tag{3.115}$$

where $\hat{Y} = a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d$

To obtain the LSE estimate of coefficients $\mathbf{A} = (a_0, a_1, a_2, \cdots, a_d)^t$, simply change the formation of matrix **X** in Eq. (3.108) to the following:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^d \end{pmatrix} \tag{3.116}$$

Note that x_i^j in Eq. (3.108) means the *j*-th dimension of sample $\mathbf{x}_i$, while x_i^j in Eq. (3.116) means *j*-th order of value x_i. Therefore, the LSE estimate of polynomial coefficients $\mathbf{A}_{LSE} = (a_0, a_1, a_2, \cdots, a_d)^t$ has the same form as Eq. (3.112).

3.2.1.3. MMSE/LSE For Nonlinear Functions

As the most general case, consider solving the following minimization problem:

$$\min_{g(\bullet) \in G_{nl}} E\left[Y - g(X)\right]^2 \tag{3.117}$$

Since we need to deal with all possible nonlinear functions, taking a derivative does not work here. Instead, we use the property of conditional expectation to solve this minimization problem. By applying Eq. (3.38) to (3.117), we get

$$\begin{aligned} E_{X,Y}\left[Y - g(X)\right]^2 &= E_X\left\{E_{Y|X}\left[\left[Y - g(X)\right]^2 \mid X = x\right]\right\} \\ &= \int_{-\infty}^{\infty} E_{Y|X}\left[\left[Y - g(X)\right]^2 \mid X = x\right] f_X(x)dx \\ &= \int_{-\infty}^{\infty} E_{Y|X}\left[\left[Y - g(x)\right]^2 \mid X = x\right] f_X(x)dx \end{aligned} \tag{3.118}$$

Since the integrand is nonnegative in Eq. (3.118), the quantity in Eq. (3.117) will be minimized at the same time the following equation is minimized.

$$\min_{g(x) \in R} E_{Y|X}\left[\left[Y - g(x)\right]^2 \mid X = x\right] \tag{3.119}$$

Since $g(x)$ is a constant in the calculation of the conditional expectation above, the MMSE estimate can be obtained in the same way as the constant functions in Section 3.2.1.1. Thus, the MMSE estimate should take the following form:

$$\hat{Y} = g_{MMSE}(X) = E_{Y|X}(Y \mid X) \tag{3.120}$$

If the value $X = x$ is observed and the value $E(Y \mid X = x)$ is used to predict *Y*, the mean squared error (MSE) is minimized and specified as follows:

$$E_{Y|X}\left[\left[Y - E_{Y|X}(Y \mid X = x)\right]^2 \mid X = x\right] = Var_{Y|X}(Y \mid X = x) \tag{3.121}$$

The overall MSE, averaged over all the possible values of *X*, is:

$$E_X\left[Y - E_{Y|X}(Y \mid X)\right]^2 = E_X\left\{E_{Y|X}\left[\left[Y - E_{Y|X}(Y \mid X)\right]^2 \mid X\right]\right\} = E_X\left[{}_{Y|X}Var(Y \mid X = x)\right] \tag{3.122}$$

It is important to distinguish between the overall MSE $E_X\left[Var_{Y|X}(Y \mid X)\right]$ and the MSE of the particular estimate when $X = x$, which is $Var_{Y|X}(Y \mid X = x)$. Before the value of *X* is observed, the expected MSE for the process of observing *X* and predicting *Y* is $E_X\left[Var_{Y|X}(Y \mid X)\right]$. On the other hand, after a particular value *x* of *X* has been observed and the prediction $E_{Y|X}(Y \mid X = x)$ has been made, the appropriate measure of MSE of the prediction is $Var_{Y|X}(Y \mid X = x)$.

In general, the form of the MMSE estimator for nonlinear functions depends on the form of the joint distribution of *X* and *Y*. There is no mathematical closed-form solution. To get the conditional expectation in Eq. (3.120), we have to perform the following integral:

$$\hat{Y}(x) = \int_{-\infty}^{\infty} y f_Y(y \mid X = x)dy \tag{3.123}$$

It is difficult to solve this integral calculation. First, different measures of *x* could determine different conditional p.d.f. for the integral. Exact information about the p.d.f. is often impossible to obtain. Second, there could be no analytic solution for the integral. Those difficulties reduce the interest of the MMSE estimation of nonlinear functions to theoretical aspects only. The same difficulties also exist for LSE estimation for nonlinear functions. Some certain classes of well-behaved nonlinear functions are typically assumed for LSE problems and numeric methods are used to obtain LSE estimate from sample data.

3.2.2. Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is the most widely used parametric estimation method, largely because of its efficiency. Suppose that a set of random samples $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ is to be drawn independently according to a discrete or continuous distribution with the p.f. or the p.d.f. $p(x \mid \mathbf{\Phi})$, where the parameter vector $\mathbf{\Phi}$ belongs to some parameter space Ω. Given an observed vector $\mathbf{x} = (x_1, \cdots, x_n)$, the *likelihood* of the set of sample data vectors $\mathbf{x}$ with respect to $\mathbf{\Phi}$ is defined as the joint p.f. or joint p.d.f. $p_n(\mathbf{x} \mid \mathbf{\Phi})$; $p_n(\mathbf{x} \mid \mathbf{\Phi})$ is also referred to as the *likelihood function*.

MLE assumes the parameters of p.d.f.'s are fixed but unknown and aims to find the set of parameters that maximizes the likelihood of generating the observed data. For example, the p.d.f. $p_n(\mathbf{x} \mid \mathbf{\Phi})$ is assumed to be a Gaussian distribution $N(\boldsymbol{\mu}, \mathbf{\Sigma})$, the components of $\mathbf{\Phi}$ will then include exactly the components of mean-vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Since $X_1, X_2, \ldots, X_n$ are independent random variables, the likelihood can be rewritten as follows:

$$p_n(\mathbf{x} \mid \mathbf{\Phi}) = \prod_{k=1}^{n} p(x_k \mid \mathbf{\Phi}) \tag{3.124}$$

The likelihood $p_n(\mathbf{x} \mid \mathbf{\Phi})$ can be viewed as the probability of generating the sample data set $\mathbf{x}$ based on parameter set $\mathbf{\Phi}$. The *maximum likelihood estimator* of $\mathbf{\Phi}$ is denoted as $\mathbf{\Phi}_{MLE}$ that maximizes the likelihood $p_n(\mathbf{x} \mid \mathbf{\Phi})$. That is,

$$\mathbf{\Phi}_{MLE} = \underset{\Phi}{\operatorname{argmax}}\; p_n(\mathbf{x} \mid \mathbf{\Phi}) \tag{3.125}$$

This estimation method is called the *maximum likelihood estimation* method and is often abbreviated as MLE. Since the logarithm function is a monotonically increasing function, the parameter set $\mathbf{\Phi}_{MLE}$ that maximizes the log-likelihood should also maximize the likelihood. If $p_n(\mathbf{x} \mid \Phi)$ is differentiable function of $\mathbf{\Phi}$, $\mathbf{\Phi}_{MLE}$ can be attained by taking the partial derivative with respect to Φ and setting it to zero. Specifically, let $\mathbf{\Phi}$ be a *k*-component parameter vector $\mathbf{\Phi} = (\Phi_1, \Phi_2, \ldots, \Phi_k)^t$ and $\nabla_{\mathbf{\Phi}}$ be the gradient operator:

$$\nabla_{\mathbf{\Phi}} = \begin{bmatrix} \dfrac{\partial}{\partial \Phi_1} \\ \vdots \\ \dfrac{\partial}{\partial \Phi_k} \end{bmatrix} \tag{3.126}$$

The log-likelihood becomes:

$$l(\mathbf{\Phi}) = \log p_n(\mathbf{x} \mid \mathbf{\Phi}) = \sum_{k=1}^{n} \log p(x_k \mid \mathbf{\Phi}) \tag{3.127}$$

and its partial derivative is:

$$\nabla_{\mathbf{\Phi}}\; l(\mathbf{\Phi}) = \sum_{k=1}^{n} \nabla_{\mathbf{\Phi}} \log p(x_k \mid \mathbf{\Phi}) \tag{3.128}$$

Thus, the maximum likelihood estimate of Φ can be obtained by solving the following set of *k* equations:

$$\nabla_{\mathbf{\Phi}}\; l(\mathbf{\Phi}) = 0 \tag{3.129}$$

Example 1

Let's take a look at the maximum likelihood estimator of a univariate Gaussian p.d.f., given as the following equation:

$$p(x\,|\,\mathbf{\Phi}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{3.130}$$

where μ and σ^2 are the mean and the variance respectively. The parameter vector $\mathbf{\Phi}$ denotes (μ,σ^2) . The log-likelihood is:

$$\begin{aligned}
&\log p_n(\mathbf{x}\,|\,\mathbf{\Phi}) = \sum_{k=1}^{n}\log p(x_k\,|\,\mathbf{\Phi}) \\
&= \sum_{k=1}^{n}\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]\right) \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2
\end{aligned} \tag{3.131}$$

and the partial derivative of the above expression is:

$$\begin{aligned}
&\frac{\partial}{\partial\mu}\log p_n(x\,|\,\mathbf{\Phi}) = \sum_{k=1}^{n}\frac{1}{\sigma^2}(x_k-\mu) \\
&\frac{\partial}{\partial\sigma^2}\log p_n(x\,|\,\mathbf{\Phi}) = -\frac{n}{2\sigma^2} + \sum_{k=1}^{n}\frac{(x_k-\mu)^2}{2\sigma^4}
\end{aligned} \tag{3.132}$$

We set the two partial differential derivatives to zero,

$$\begin{aligned}
&\sum_{k=1}^{n}\frac{1}{\sigma^2}(x_k-\mu) = 0 \\
&-\frac{n}{\sigma^2} + \sum_{k=1}^{n}\frac{(x_k-\mu)^2}{\sigma^4} = 0
\end{aligned} \tag{3.133}$$

The maximum likelihood estimates for μ and σ^2 are obtained by solving the above equations:

$$\begin{aligned}
\mu_{MLE} &= \frac{1}{n}\sum_{k=1}^{n}x_k = E(x) \\
\sigma^2_{MLE} &= \frac{1}{n}\sum_{k=1}^{n}(x_k-\mu_{MLE})^2 = E\left[(x-\mu_{MLE})^2\right]
\end{aligned} \tag{3.134}$$

Equation (3.134) indicates that the maximum likelihood estimation for mean and variance is just the sample mean and variance.

Example 2

For the multivariate Gaussian p.d.f. $p(\mathbf{x})$

$$p(\mathbf{x} \mid \mathbf{\Phi}) = \frac{1}{(2\pi)^{d/2} \left|\mathbf{\Sigma}\right|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right] \tag{3.135}$$

The maximum likelihood estimates of $\mathbf{m}$ and Σ can be obtained by a similar procedure.

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{MLE} &= \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k \\ \hat{\mathbf{\Sigma}}_{MLE} &= \frac{1}{n} \sum_{k=1}^{n} (\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{MLE})(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{MLE})^t = E\left[(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{MLE})(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{MLE})^t \right] \end{aligned} \tag{3.136}$$

Once again, the maximum likelihood estimation for mean vector and co-variance matrix is the sample mean vector and sample covariance matrix.

In some situations, a maximum likelihood estimation of $\mathbf{\Phi}$ may not exist, or the maximum likelihood estimator may not be uniquely defined, i.e., there may be more than one MLE of $\mathbf{\Phi}$ for a specific set of sample values. Fortunately, according to Fisher's theorem, for most practical problems with a well-behaved family of distributions, the MLE exists and is uniquely defined [4, 25, 26].

In fact, the maximum likelihood estimator can be proven to be sound under certain conditions. As mentioned before, the estimator $\theta(\mathbf{X})$ is a function of the vector of random variables $\mathbf{X}$ that represent the sample data. $\theta(\mathbf{X})$ itself is also a random variable, with a distribution determined by joint distributions of $\mathbf{X}$. Let $\tilde{\mathbf{\Phi}}$ be the parameter vector of true distribution $p(x \mid \mathbf{\Phi})$ from which the samples are drawn. If the following three conditions hold:

1. The sample $\mathbf{x}$ is a drawn from the assumed family of distribution,
2. The family of distributions is well behaved,
3. The sample $\mathbf{x}$ is large enough,

then maximum likelihood estimator, $\mathbf{\Phi}_{MLE}$, has a Gaussian distribution with a mean $\tilde{\mathbf{\Phi}}$ and a variance of the form $1/nB_{\mathbf{x}}^2$ [26], where *n* is the size of sample and $B_{\mathbf{x}}$ is the *Fisher information*, which is determined solely by $\tilde{\mathbf{\Phi}}$ and $\mathbf{x}$. An estimator is said to be *consistent*, iff the estimate will converge to the true distribution when there is infinite number of training samples.

$$\lim_{n->\infty} \mathbf{\Phi}_{MLE} = \tilde{\mathbf{\Phi}} \tag{3.137}$$

$\mathbf{\Phi}_{MLE}$ is a consistent estimator based on the analysis above. In addition, it can be shown that no consistent estimator has a lower variance than $\mathbf{\Phi}_{MLE}$. In other words, no estimator provides a closer estimate of the true parameters than the maximum likelihood estimator.

3.2.3. Bayesian Estimation and MAP Estimation

Bayesian estimation has a different philosophy than maximum likelihood estimation. While MLE assumes that the parameter Φ [3] is fixed but unknown, Bayesian estimation assumes that the parameter Φ itself is a random variable with a prior distribution $p(\Phi)$. Suppose we observe a sequence of random samples $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, which are i.i.d. with a p.d.f. $p(x \mid \Phi)$. According to Bayes' rule, we have the posterior distribution of Φ as:

$$p(\Phi \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \Phi)p(\Phi)}{p(\mathbf{x})} \propto p(\mathbf{x} \mid \Phi)p(\Phi) \tag{3.138}$$

In Eq. (3.138), we dropped the denominator $p(\mathbf{x})$ here because it is independent of the parameter Φ. The distribution in Eq. (3.138) is called the posterior distribution of Φ because it is the distribution of Φ after we observed the values of random variables $X_1, X_2, \ldots, X_n$.

3.2.3.1. Prior and Posterior Distributions

For mathematical tractability, conjugate priors are often used in Bayesian estimation. Suppose a random sample is taken of a known distribution with p.d.f. $p(\mathbf{x} \mid \Phi)$. A conjugate prior for the random variable (or vector) is defined as the prior distribution for the parameters of the probability density function of the random variable (or vector), such that the class-conditional p.d.f. $p(\mathbf{x} \mid \Phi)$, the posterior distribution $p(\Phi \mid \mathbf{x})$, and the prior distribution $p(\Phi)$ belong to the same distribution family. For example, it is well known that the conjugate prior for the mean of a Gaussian p.d.f. is also a Gaussian p.d.f. [4]. Now, let's derive such a posterior distribution $p(\Phi \mid \mathbf{x})$ from the widely used Gaussian conjugate prior.

Example

Suppose $X_1, X_2, \ldots, X_n$ are drawn from a Gaussian distribution for which the mean Φ is a random variable and the variance σ^2 is known. The likelihood function $p(\mathbf{x} \mid \Phi)$ can be written as:

[3] For simplicity, we assume the parameter Φ is a scalar instead of a vector here. However, the extension to a parameter vector $\mathbf{\Phi}$ can be derived according to a similar procedure.

$$p(\mathbf{x}\,|\,\Phi)=\frac{1}{(2\pi)^{n/2}\sigma^{n}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\Phi}{\sigma}\right)^2\right]\propto\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\Phi}{\sigma}\right)^2\right] \tag{3.139}$$

To further simply Eq. (3.139), we could use Eq. (3.140)

$$\sum_{i=1}^{n}(x_i-\Phi)^2=n(\Phi-\overline{x}_n)^2+\sum_{i=1}^{n}(x_i-\overline{x}_n)^2 \tag{3.140}$$

where $\overline{x}_n=\frac{1}{n}\sum_{i=1}^{n}x_i=$ the sample mean of $\mathbf{x}=\{x_1,x_2,\ldots,x_n\}$.

Let's rewrite $p(\mathbf{x}\,|\,\Phi)$ in Eq. (3.139) into Eq. (3.141):

$$p(\mathbf{x}\,|\,\Phi)\propto\exp\left[-\frac{n}{2\sigma^2}(\Phi-\overline{x}_n)^2\right]\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right] \tag{3.141}$$

Now supposed the prior distribution of Φ is also a Gaussian distribution with mean μ and variance ν^2, i.e., the prior distribution $p(\Phi)$ is given as follows:

$$p(\Phi)=\frac{1}{(2\pi)^{1/2}\nu}\exp\left[-\frac{1}{2}\left(\frac{\Phi-\mu}{\nu}\right)^2\right]\propto\exp\left[-\frac{1}{2}\left(\frac{\Phi-\mu}{\nu}\right)^2\right] \tag{3.142}$$

By combining Eqs. (3.141) and (3.142) while dropping the second term in Eq. (3.141) we could attain the posterior p.d.f. $p(\Phi\,|\,\mathbf{x})$ in the following equation:

$$p(\Phi\,|\,\mathbf{x})\propto\exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\Phi-\overline{x}_n)^2+\frac{1}{\nu^2}(\Phi-\mu)^2\right]\right\} \tag{3.143}$$

Now if we define ρ and τ as follows:

$$\rho=\frac{\sigma^2\mu+n\nu^2\overline{x}_n}{\sigma^2+n\nu^2} \tag{3.144}$$

$$\tau^2=\frac{\sigma^2\nu^2}{\sigma^2+n\nu^2} \tag{3.145}$$

We can rewrite Eq. (3.143) can be rewritten as:

$$p(\Phi\,|\,\mathbf{x})\propto\exp\left\{-\frac{1}{2}\left[\frac{1}{\tau^2}(\Phi-\rho)^2+\frac{n}{\sigma^2+n\nu^2}(\overline{x}_n-\mu)^2\right]\right\} \tag{3.146}$$

Since the second term in Eq. (3.146) does not depend on Φ, it can be absorbed in the constant factor. Finally, we have the posterior p.d.f. in the following form:

$$p(\Phi \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi}\tau} \exp\left[\frac{-1}{2\tau^2} (\Phi - \rho)^2 \right] \tag{3.147}$$

Equation (3.147) shows that the posterior p.d.f. $p(\Phi \mid \mathbf{x})$ is a Gaussian distribution with mean ρ and variance τ^2 as defined in Eqs. (3.144) and (3.145). The Gaussian prior distribution defined in Eq. (3.142) is a conjurgate prior.

3.2.3.2. General Bayesian Estimation

The foremost requirement of a good estimator θ is that it can yield an estimate of Φ ($\theta(\mathbf{X})$) which is close to the real value Φ. In other words, a good estimator is one for which it is highly probable that the error $\theta(\mathbf{X}) - \Phi$ is close to 0. In general, we can define a loss function[4] $R(\Phi, \bar{\Phi})$. It measures the loss or cost associated with the fact that the true value of the parameter is Φ while the estimate is $\bar{\Phi}$. When only the prior distribution $p(\Phi)$ is available and no sample data has been observed, if we choose one particular estimate $\bar{\Phi}$, the expected loss is:

$$E\left[R(\Phi, \bar{\Phi}) \right] = \int R(\Phi, \bar{\Phi}) p(\Phi) d\Phi \tag{3.148}$$

The fact that we could derive posterior distribution from the likelihood function and the prior distribution [as shown in the derivation of Eq. (3.147)] is very important here because it allows us to compute the expected posterior loss after sample vector $\mathbf{x}$ is observed. The expected posterior loss associated with estimate $\bar{\Phi}$ is:

$$E\left[R(\Phi, \bar{\Phi}) \mid \mathbf{x} \right] = \int R(\Phi, \bar{\Phi}) p(\Phi \mid \mathbf{x}) d\Phi \tag{3.149}$$

The Bayesian estimator of Φ is defined as the estimator that attains minimum Bayes risk, that is, minimizes the expected posterior loss function (3.149). Formally, the Bayesian estimator is chosen according to:

$$\theta_{Bayes}(\mathbf{x}) = \underset{\theta}{\operatorname{argmin}}\ E\left[R(\Phi, \theta(\mathbf{x})) \mid \mathbf{x} \right] \tag{3.150}$$

The Bayesian estimator of Φ is the estimator θ_{Bayes} for which Eq. (3.150) is satisfied for every possible value of $\mathbf{x}$ of random vector $\mathbf{X}$. Therefore, the form of the Bayesian estimator θ_{Bayes} should depend only on the loss function and the prior distribution, but not the sample value $\mathbf{x}$.

[4] The Bayesian estimation and loss function are based on Bayes' decision theory, which will be described in detail in Chapter 4.

One of the most common loss functions used in statistical estimation is the mean squared error function [20]. The mean squared error function for Bayesian estimation should have the following form:

$$R(\Phi,\theta(\mathbf{x}))=(\Phi-\theta(\mathbf{x}))^2 \tag{3.151}$$

In order to find the Bayesian estimator, we are seeking θ_{Bayes} to minimize the expected posterior loss function:

$$E\left[R(\Phi,\theta(\mathbf{x}))\,|\,\mathbf{x}\right]=E\left[(\Phi-\theta(\mathbf{x}))^2\,|\,\mathbf{x}\right]=E(\Phi^2\,|\,\mathbf{x})-2\theta(\mathbf{x})E(\Phi\,|\,\mathbf{x})-\theta(\mathbf{x})^2 \tag{3.152}$$

The minimum value of this function can be obtained by taking the partial derivative of Eq. (3.152) with respect to $\theta(\mathbf{x})$. Since the above equation is simply a quadratic function of $\theta(\mathbf{x})$, it can be shown that the minimum loss can be achieved when θ_{Bayes} is chosen based on the following equation:

$$\theta_{Bayes}(\mathbf{x})=E(\Phi\,|\,\mathbf{x}) \tag{3.153}$$

Equation (3.153) translates into the fact that the Bayesian estimate of the parameter Φ for mean squared error function is equal to the mean of the posterior distribution of Φ. In the following section, we discuss another popular loss function (MAP estimation) that also generates the same estimate for certain distribution functions.

3.2.3.3. MAP Estimation

One intuitive interpretation of Eq. (3.138) is that a prior p.d.f. $p(\Phi)$ represents the relative likelihood before the values of $X_1, X_2, \ldots, X_n$ have been observed; while the posterior p.d.f. $p(\Phi\,|\,\mathbf{x})$ represents the relative likelihood after the values of $X_1, X_2, \ldots, X_n$ have been observed. Therefore, choosing an estimate $\bar{\Phi}$ that maximizes posterior probability is consistent with out intuition. This estimator is in fact the *maximum posterior probability* (MAP) estimator and is the most popular Bayesian estimator.

The loss function associated with the MAP estimator is the so-called uniform loss function [20]:

$$R(\Phi,\theta(\mathbf{x}))=\begin{cases}0, & \text{if } |\theta(\mathbf{x})-\Phi|\le\Delta \\ 1, & \text{if } |\theta(\mathbf{x})-\Phi|>\Delta\end{cases} \quad \text{where } \Delta>0 \tag{3.154}$$

Now let’s see how this uniform loss function results in MAP estimation. Based on loss function defined above, the expected posterior loss function is:

$$\begin{aligned}E(R(\Phi,\theta(\mathbf{x}))\,|\,\mathbf{x})&=P(|\theta(\mathbf{x})-\Phi|>\Delta\,|\,\mathbf{x})\\&=1-P(|\theta(\mathbf{x})-\Phi|\le\Delta\,|\,\mathbf{x})=1-\int_{\theta(\mathbf{x})-\Delta}^{\theta(\mathbf{x})+\Delta}p(\Phi\,|\,\mathbf{x})\end{aligned} \tag{3.155}$$

The quantity in Eq. (3.155) is minimized by maximizing the shaded area under $p(\Phi \mid \mathbf{x})$ over the interval $\left[\theta(\mathbf{x})-\Delta, \theta(\mathbf{x})+\Delta\right]$ in Figure 3.16. If $p(\Phi \mid \mathbf{x})$ is a smooth curve and Δ is small enough, the shaded area can be computed roughly as:

$$\int_{\theta(\mathbf{x})-\Delta}^{\theta(\mathbf{x})+\Delta} p(\Phi \mid \mathbf{x}) \cong 2\Delta p(\Phi \mid \mathbf{x}) \Big|_{\Phi=\theta(\mathbf{x})} \tag{3.156}$$

Thus, the shaded area can be approximately maximized by choosing $\theta(\mathbf{x})$ to be the maximum point of $p(\Phi \mid \mathbf{x})$. This concludes our proof the using the error function in Eq. (3.154) indeed will generate MAP estimator.

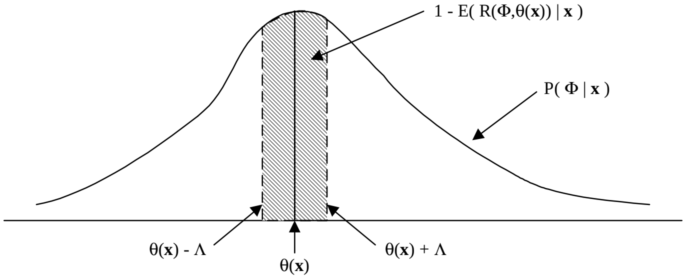

Figure 3.16 Illustration of finding the minimum expected posterior loss function for MAP estimation [20].

MAP estimation is to find the parameter estimate Φ_{MAP} or estimator $\theta_{MAP}(\mathbf{x})$ that maximizes the posterior probability,

$$\Phi_{MAP} = \theta_{MAP}(\mathbf{x}) = \underset{\Phi}{\operatorname{argmax}}\ p(\Phi \mid \mathbf{x}) = \underset{\Phi}{\operatorname{argmax}}\ p(\mathbf{x} \mid \Phi) p(\Phi) \tag{3.157}$$

Φ_{MAP} can also be specified in the logarithm form as follows:

$$\Phi_{MAP} = \underset{\Phi}{\operatorname{argmax}}\ \log p(\mathbf{x} \mid \Phi) + \log p(\Phi) \tag{3.158}$$

Φ_{MAP} can be attained by solving the following partial differential equation:

$$\frac{\partial \log p(\mathbf{x} \mid \Phi)}{\partial \Phi} + \frac{\partial \log p(\Phi)}{\partial \Phi} = 0 \tag{3.159}$$

Thus the MAP equation for finding Φ_{MAP} can be established.

$$\frac{\partial \log p(\mathbf{x} \mid \Phi)}{\partial \Phi} \Big|_{\Phi=\Phi_{MAP}} = \frac{-\partial \log p(\Phi)}{\partial \Phi} \Big|_{\Phi=\Phi_{MAP}} \tag{3.160}$$

There are interesting relationships between MAP estimation and MLE estimation. The prior distribution is viewed as the knowledge of the statistics of the parameters of interest before any sample data is observed. For the case of MLE, the parameter is assumed to be fixed but unknown. That is, there is no preference (knowledge) of what the values of parameters should be. The prior distribution $p(\Phi)$ can only be set to constant for the entire parameter space, and this type of prior information is often referred to as *non-informative prior* or *uniform prior*. By substituting $p(\Phi)$ with a uniform distribution in Eq. (3.157), MAP estimation is identical to MLE. In this case, the parameter estimation is solely determined by the observed data. A sufficient amount of training data is often a requirement for MLE. On the other hand, when the size of the training data is limited, the use of the prior density becomes valuable. If some prior knowledge of the distribution of the parameters can be obtained, the MAP estimation provides a way of incorporating prior information in the parameter learning process.

Example

Now, let's formulate MAP estimation for Gaussian densities. As described in Section 3.2.3.1, the conjugate prior distribution for a Gaussian density is also a Gaussian distribution. Similarly, we assumed random variables $X_1, X_2, \ldots, X_n$ drawn from a Gaussian distribution for which the mean Φ is unknown and the variance σ^2 is known, while the conjugate prior distribution of Φ is a Gaussian distribution with mean μ and variance ν^2. It is shown in Section 3.2.3.1 that the posterior p.d.f. can be formulated as in Eq. (3.147). The MAP estimation for Φ can be solved by taking the derivative of Eq. (3.147) with respect to Φ:

$$\Phi_{MAP} = \rho = \frac{\sigma^2 \mu + n\nu^2 \overline{x}_n}{\sigma^2 + n\nu^2} \qquad (3.161)$$

where n is the total number of training samples and $\overline{x}_n$ the sample mean.

The MAP estimate of the mean Φ is a weighted average of the sample mean $\overline{x}_n$ and the prior mean. When n is zero (when there is no training data at all), the MAP estimate is simply the prior mean μ. On the other hand, when n is large ($n \to \infty$), the MAP estimate will converge to the maximum likelihood estimate. This phenomenon is consistent with our intuition and is often referred to as *asymptotic equivalence* or *asymptotic convergence*. Therefore, in practice, the difference between MAP estimation and MLE is often insignificant when a large amount of training data is available. When the prior variance ν^2 is very large (e.g., $\nu^2 >> \sigma^2 / n$), the MAP estimate will converge to the ML estimate because a very large ν^2 translated into a non-informative prior.

It is important to note that the requirement of learning prior distribution for MAP estimation is critical. In some cases, the prior distribution is very difficult to estimate and MLE is still an attractive estimation method. As mentioned before, the MAP estimation framework is particularly useful for dealing with sparse data, such as parameter adaptation. For example, in speaker adaptation, the speaker-independent (or multiple speakers) database can be used to first estimate the prior distribution [9]. The model parameters are adapted to a target speaker through a MAP framework by using limited speaker-specific training data as discussed in Chapter 9.

3.3. SIGNIFICANCE TESTING

Significance testing is one of the most important theories and methods of statistical *inference*. A problem of statistical inference, or, more simply, a statistics problem, is one in which data that have been generated in accordance with some unknown probability distribution must be analyzed, and some type of inference about the unknown distribution must be made. Hundreds of test procedures have developed in statistics for various kinds of hypotheses testing. We focus only on tests that are used in spoken language systems.

The selection of appropriate models for the data or systems is essential for spoken language systems. When the distribution of certain sample data is unknown, it is usually appropriate to make some assumptions about the distribution of the data with a distribution function whose properties are well known. For example, people often use Gaussian distributions to model the distribution of background noise in spoken language systems. One important issue is how good our assumptions are, and what the appropriate values of the parameters for the distributions are, even when we can use the methods in Section 3.2 to estimate parameters from sample data. Statistical tests are often applied to determine if the distribution with specific parameters is appropriate to model the sample data. In this section, we describe the most popular testing method for the goodness of distribution fitting – the χ^2 goodness-of-fit test.

Another important type of statistical tests is designed to evaluate the excellence of two different methods or algorithms for the same tasks when there is uncertainty regarding the results. To assure that the two systems are evaluated on the same or similar conditions, experimenters often carefully choose similar or even the exactly same data sets for testing. This is why we refer to this type of statistical test as a *paired observations* test. In both speech recognition and speech synthesis, the paired observations test is a very important tool for interpreting the comparison results.

3.3.1. Level of Significance

We now consider statistical problems involving a parameter ϕ whose value is unknown but must lie in a certain parameter space Ω. In statistical tests, we let H_0 denote the hypothesis that $\phi \in \Omega_0$ and let H_1 denote the hypothesis that $\phi \in \Omega_1$. The subsets Ω_0 and Ω_1 are disjoint and $\Omega_0 \cup \Omega_1 = \Omega$, so exactly one of the hypotheses H_0 and H_1 must be true. We must now decide whether to accept H_0 or H_1 by observing a random sample $X_1, \cdots, X_n$ drawn from a distribution involving the unknown parameter ϕ. A problem like this is called hypotheses testing. A procedure for deciding whether to accept H_0 or H_1 is called a *test procedure* or simply a *test.* The hypothesis H_0 is often referred to as the *null hypothesis* and the hypothesis H_1 as the *alternative hypothesis*. Since there are only two possible decisions, accepting H_0 is equivalent to rejecting H_1 and rejecting H_0 is equivalent to accepting H_1. Therefore, in testing hypotheses, we often use the terms *accepting or rejecting the null hypothesis* H_0 as the only decision choices.

Usually we are presented with a random sample $\mathbf{X} = (X_1, \cdots, X_n)$ to help us in making the test decision. Let *S* denote the sample space of *n*-dimensional random vector **X**. The testing procedure is equivalent to partitioning the sample space *S* into two subsets. One subset specifies the values of **X** for which one will accept H_0 and the other subset specifies the values of **X** for which one will reject H_0. The second subset is called the *critical region* and is often denoted as *C*.

Since there is uncertainty associated with the test decision, for each value of $\phi \in \Omega$, we are interested in the probability $\rho(\phi)$ that the testing procedure rejects H_0. The function $\rho(\phi)$ is called the *power function* of the test and can be specified as follows:

$$\rho(\phi) = P(\mathbf{X} \in C \mid \phi) \qquad (3.162)$$

For $\phi \in \Omega_0$, the decision to reject H_0 is incorrect. Therefore, if $\phi \in \Omega_0$, $\rho(\phi)$ is the probability that the statistician will make an incorrect decision (false rejection). In statistical tests, an upper bound α_0 $(0 < \alpha_0 < 1)$ is specified, and we only consider tests for which $\rho(\phi) \le \alpha_0$ for every value of $\phi \in \Omega_0$. The upper bound α_0 is called the *level of significance*. The smaller α_0 is, the less likely it is that the test procedure will reject H_0. Since α_0 specifies the upper bound for false rejection, once a hypothesis is rejected by the test procedure, we can be $(1-\alpha_0)$ confident the decision is correct. In most applications, α_0 is set to be 0.05 and the test is said to be carried out at the 0.05 level of significance or 0.95 level of confidence.

We define the size α of a given test as the maximum probability, among all the values of ϕ which satisfy the null hypothesis, of making an incorrect decision.

$$\alpha = \max_{\theta \in \Omega_0} \rho(\phi) \qquad (3.163)$$

Once we obtain the value of α, the test procedure is straightforward. First, the statistician specifies a certain level of significance α_0 in a given problem of testing hypotheses, then he or she rejects the null hypothesis if the size α is such that $\alpha \le \alpha_0$.

The size α of a given test is also called the *tail area* or the *p-value* corresponding to the observed value of data sample **X** because it corresponds to tail area of the distribution. The hypothesis will be rejected if the level of significance α_0 is such that $\alpha_0 > \alpha$ and should be accepted for any value of $\alpha_0 < \alpha$. Alternatively, we can say the observed value of

X is *just significant* at the level of significance α without using the level of significance α_0. Therefore, if we had found that the observed value of one data sample **X** was just significant at the level of 0.0001, while the other observed value of data sample **Y** was just significant at the level of 0.001, then we can conclude the sample **X** provides much stronger evidence against H_0. In statistics, an observed value of one data sample **X** is generally said to be statistically significant if the corresponding tail area is smaller than the traditional value 0.05. For case requiring more significance (confidence), 0.01 can be used.

A statistically significant observed data sample **X** that provides strong evidence against H_0 does not necessary provide strong evidence that the actual value of ϕ is significantly far away from parameter set Ω_0. This situation can arise, particularly when the size of random data sample is large, because a test with larger sample size will in general reject hypotheses with more confidence, unless the hypothesis is indeed the true one.

3.3.2. Normal Test (*Z*-Test)

Suppose we need to find whether a coin is fair or not. Let *p* be the probability of the head. The hypotheses to be tested are as follows:

$H_0 \,:\, p = \frac{1}{2}$

$H_1 \,:\, p \neq \frac{1}{2}$

We assume that a random sample size *n* is taken, and let random variable *M* denote the number of times we observe heads as the result. The random variable *M* has a binomial distribution $B(n, \frac{1}{2})$. Because of the shape of binomial distribution, *M* can lie on either side of the mean. This is why it is called a typical *two-tailed test*. The tail area or *p*-value for the observed value *k* can be computed as:

$$p = \begin{cases} 2P(k \le M \le n) & \text{for } k > n/2 \\ 2P(0 \le M \le k) & \text{for } k < n/2 \\ 1.0 & \text{for } k = n/2 \end{cases} \tag{3.164}$$

The *p*-value in Eq. (3.164) can be computed directly using the binomial distribution. The test procedure will reject H_0 when *p* is less than the significance level α_0.

In many situations, the *p*-value for the distribution of data sample **X** is difficult to obtain due to the complexity of the distribution. Fortunately, if some statistic *Z* of the data sample **X** has some well-known distribution, the test can then be done in the *Z* domain instead. If *n* is large enough ($n > 50$), a *normal test* (or *Z-test*) can be used to approximate a binomial probability. Under H_0, the mean and variance for *M* are $E(M) = n/2$ and $Var(M) = n/4$. The new random variable *Z* is defined as,

$$Z = \frac{|M - n/2| - 1/2}{\sqrt{n/4}} \tag{3.165}$$

which can be approximated as standard Gaussian distribution $N(0,1)$ under H_0. The p-value can now be computed as $p = 2P(Z \geq z)$ where z is the realized value of Z after M is observed. Thus, H_0 is rejected if $p < \alpha_0$, where α_0 is the level of significance.

3.3.3. χ^2 Goodness-of-Fit Test

The normal test (Z-test) can be extended to test the hypothesis that a given set of data came from a certain distribution with all parameters specified. First let's look at the case of discrete distribution fitting.

Suppose that a large population consists of items of k different types and let p_i be the probability that a random selected item belongs to type i. Now, let $q_1, \ldots, q_k$ be a set of specific numbers satisfying the probabilistic constraint ($q_i \geq 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^{k} q_i = 1$). Finally, suppose that the following hypotheses are to be tested:

H_0 : $p_i = q_i$ for $i = 1, \ldots, k$

H_1 : $p_i \neq q_i$ for at least one value of i

Assume that a random sample of size n is to be taken from the given population. For $i = 1, \ldots, k$, let N_i denote the number of observations in the random sample which are of type i. Here, $N_1, \ldots, N_k$ are nonnegative numbers and $\sum_{i=1}^{k} N_i = n$. Random variables $N_1, \ldots, N_k$ have a multinomial distribution. Since the p-value for the multinomial distribution is hard to obtain, instead we use another statistic about $N_1, \ldots, N_k$. When H_0 is true, the expected number of observations of type i is nq_i. In other words, the difference between the actual number of observations N_i and the expected number nq_i should be small when H_0 is true. It seems reasonable to base the test on the differences $N_i - nq_i$ and to reject H_0 when the differences are large. It can be proved [14] that the following random variable λ

$$\lambda = \sum_{i=1}^{k} \frac{(N_i - nq_i)^2}{nq_i} \tag{3.166}$$

converges to the χ^2 distribution with $k-1$ degrees of freedom as the sample size $n \to \infty$.

A χ^2 test of goodness-of-fit can be carried out in the following way. Once a level of significance α_0 is specified, we can use the following p-value function to find critical point c[5]:

$$P(\lambda > c) = 1 - F_{\chi^2}(x = c) = \alpha_0 \tag{3.167}$$

[5] Since χ^2 pdf is a monotonic function, the test is a one-tail test. Thus, we only need to calculate one tail area.

where $F_{\chi^2}(x)$ is the distribution function for χ^2 distribution. The test procedure simply rejects H_0 when the realized value λ is such that $\lambda > c$. Empirical results show that the χ^2 distribution will be a good approximation to the actual distribution of λ as long as the value of each expectation nq_i is not too small (≥ 5). The approximation should still be satisfactory if $nq_i \geq 1.5$ for $i = 1, \ldots, k$.

For continuous distributions, a modified χ^2 goodness-of-fit test procedure can be applied. Suppose that we would like to hypothesize a null hypothesis H_0 in which continuous random sample data $X_1, \ldots, X_k$ are drawn from a certain continuous distribution with all parameters specified or estimated. Also, suppose the observed values of random sample $x_1, \ldots, x_k$ are bounded within interval Ω. First, we divide the range of the hypothesized distribution into *m* subintervals within interval Ω such that the expected number of values, say E_i, in each interval is at least 5. For $i = 1, \ldots, k$, we let N_i denote the number of observations in the *i*th subintervals. As in Eq. (3.166), one can prove that the following random variable λ

$$\lambda = \sum_{i=1}^{m} \frac{(N_i - E_i)^2}{E_i} \tag{3.168}$$

converges to the χ^2 distribution with $m - k - 1$ degrees of freedom as the sample size $n \to \infty$, where *k* is the number of parameters that must be estimated from the sample data in order to calculate the expected number of values, E_i. Once the χ^2 distribution is established, the same procedure can be used to find the critical *c* in Eq. (3.167) to make test decision.

Example

Suppose we are given a random variable *X* of sample size 100 points and we want to determine whether we can reject the following hypothesis:

$$H_0 : X \sim N(0,1) \tag{3.169}$$

To perform χ^2 goodness-of-fit test, we first divide the range of *X* into 10 subintervals. The corresponding probability falling in each subinterval, the expected number of points falling in each subinterval and the actual number of points falling in each subintervals [10] are illustrated in Table 3.1.

Table 3.1 The probability falling in each subinterval of an *N*(0,1), and 100 sample points, the expected number of points falling in each subinterval, and the actual number of points falling in each subintervals [10].

Subinterval I_i	$P(X \in I_i)$	$E_i = 100P(X \in I_i)$	N_i
[-∞, -1.6]	0.0548	5.48	2

[-1.6, -1.2]	0.0603	6.03	9
[-1.2, -0.8]	0.0968	9.68	6
[-0.8, -0.4]	0.1327	13.27	11
[-0.4, 0.0]	0.1554	15.54	19
[0.0, 0.4]	0.1554	15.54	25
[0.4, 0.8]	0.1327	13.27	17
[0.8, 1.2]	0.0968	9.68	2
[1.2, 1.6]	0.0603	6.03	6
[-1.6, ∞]	0.0548	5.48	3

The value for λ can then be calculated as follows:

$$\lambda = \sum_{i=1}^{m} \frac{(N_i - E_i)^2}{E_i} = 18.286$$

Since λ can be approximated as a χ^2 distribution with $m-k-1=10-0-1=9$ degrees of freedom, the critical point *c* at the 0.05 level of significance is calculated[6] to be 16.919 according to Eq. (3.167). Thus, we should reject the hypothesis H_0 because the calculated λ is greater than the critical point *c*.

The χ^2 goodness-of-fit test at the 0.05 significance level is in general used to determine when a hypothesized distribution is not an adequate distribution to use. To accept the distribution as a good fit, one needs to make sure the hypothesized distribution cannot be rejected at the 0.4 to 0.5 level-of-significance. The alternative is to use the χ^2 goodness-of-fit test for a number of potential distributions and select the one with smallest calculated χ^2.

When all the parameters are specified (instead of estimated), the Kolmogorov-Smirnov test [5] can also be used for the goodness-of-fit test. The Kolmogorov-Smirnov test in general is a more powerful test procedure when the sample size is relatively small.

3.3.4. Matched-Pairs Test

In this section, we discuss experiments in which two different methods (or systems) are to be compared to learn which one is better. To assure the two methods are evaluated under similar conditions, two closely resemble data samples or ideally the same data sample should be used to evaluate both methods. This type of hypotheses test is called *matched-paired* test [5].

[6] In general, we use cumulative distribution function table to find the point with specific desired cumulative probability for complicated distributions, like χ^2 distribution.

3.3.4.1. The Sign Test

For $i=1,\ldots,n$, let p_i denote the probability that method *A* is better than method *B* when testing on the *i*th paired data sample. We shall assume that the probability p_i has the same value *p* for each of the *n* pairs. Suppose we wish to test the null hypothesis that method *A* is no better than method *B*. That is, the hypotheses to be tested have the following form:

$H_0\ :\ p \le ½$

$H_1\ :\ p > ½$

Suppose that, for each pair of data samples, either one method or the other will appear to be better, and the two methods cannot tie. Under these assumptions, the *n* pairs represent *n* Bernoulli trials, for each of which there is probability *p* that method *A* yields better performance. Thus the number of pairs *M* in which method *A* yields better performance will have a binomial distribution $B(n,p)$. For the simple sign test where one needs to decide which method is better, *p* will be set to $1/2$. Hence a reasonable procedure is to reject H_0 if $M>c$, where *c* is a critical point. This procedure is called a signed test. The critical point can be found according to.

$$P(M>c)=1-F_B(x=c)=\alpha_0 \qquad (3.170)$$

where $F_B(x)$ is the distribution for binomial distribution. Thus, for observed value $M>c$, we will reject H_0.

3.3.4.2. Magnitude-Difference Test

The only information that the sign test utilizes from each pair of data samples, is the sign of the difference between two performances. To do a sign test, one does not need to obtain a numeric measurement of the magnitude of the difference between the two performances. However, if the measurement of magnitude of the difference for each pair is available, a test procedure based on the relative magnitudes of the differences can be used [11].

We assume now that the performance of each method can be measured for any data samples. For $i=1,\ldots,n$, let A_i denote the performance of the method *A* on the *i*th pair of data samples and B_i denote the performance of the method *B* on the *i*th pair of data sample. Moreover, we shall let $D_i=A_i-B_i$. Since $D_1,\ldots,D_n$ are generated on *n* different pairs of data samples, they should be independent random variables. We also assume that $D_1,\ldots,D_n$ have the same distribution. Suppose now we are interested in testing the null hypothesis that method *A* and method *B* have on the average the same performance on the n pairs of data samples.

Let μ_D be the mean of D_i. The MLE estimate of μ_D is:

$$\mu_D=\sum_{i=1}^{n}\frac{D_i}{n} \qquad (3.171)$$

The test hypotheses are:

$H_0 \; : \; \mu_D = 0$

$H_1 \; : \; \mu_D \neq 0$

The MLE estimate of the variance of D_i is

$$\sigma_D^2 = \frac{1}{n}\sum_{i=1}^{n}(D_i - \mu_D)^2 \tag{3.172}$$

We define a new random variable *Z* as follows:

$$Z = \frac{\mu_D}{\sigma_D / \sqrt{n}} \tag{3.173}$$

If *n* is large enough (> 50), *Z* is proved to have a standard Gaussian distribution $N(0,1)$. The normal test procedure described in Section 3.3.2 can be used to test H_0. This type of matched-paired tests usually depends on having enough pairs of data samples for the assumption that *Z* can be approximated with a Gaussian distribution. It also requires enough data samples to estimate the mean and variance of the D_i.

3.4. INFORMATION THEORY

Transmission of information is a general definition of what we call communication. Claude Shannon's classic paper of 1948 gave birth to a new field in information theory that has become the cornerstone for coding and digital communication. In the paper titled *A Mathematical Theory of Communication,* he wrote:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

Information theory is a mathematical framework of approaching a large class of problems related to encoding, transmission, and decoding information in a systematic and disciplined way. Since speech is a form of communication, information theory has served as the underlying mathematical foundation for spoken language processing.

3.4.1. Entropy

Three interpretations can be used to describe the quantity of *information*: (1) the amount of uncertainty before seeing an event, (2) the amount of surprise when seeing an event, and (3) the amount of information after seeing an event. Although these three interpretations seem slightly different, they are virtually the same under the framework of information theory.

According to information theory, the information derivable from outcome x_i depends on its probability. If the probability $P(x_i)$ is small, we can derive a large degree of information, because the outcome that it has occurred is very rare. On the other hand, if the probability is large, the information derived will be small, because the outcome is well expected. Thus, the amount of information is defined as follows:

$$I(x_i) = \log \frac{1}{P(x_i)} \tag{3.174}$$

The reason to use a logarithm can be interpreted as follows. The information for two independent events to occur (where the joint probability is the multiplication of both individual probabilities) can be simply carried out by the addition of the individual information of each event. When the logarithm base is 2, the unit of information is called a *bit*. This means that one bit of information is required to specify the outcome. In this probabilistic framework, the amount of information represents uncertainty. Suppose X is a discrete random variable taking value x_i (referred to as a symbol) from a finite or countable infinite sample space $S = \{x_1, x_2, \ldots, x_i, \ldots\}$ (referred to as an alphabet). The symbol x_i is produced from an information source with alphabet S, according to the probability distribution of the random variable X. One of the most important properties of an information source is the entropy $H(S)$ of the random variable X, defined as the average amount of information (expected information):

$$H(X) = E[I(X)] = \sum_S P(x_i) I(x_i) = \sum_S P(x_i) \log \frac{1}{P(x_i)} = E\left[-\log P(X)\right] \tag{3.175}$$

This entropy $H(X)$ is the amount of information required to specify what kind of symbol has occurred on average. It is also the averaged uncertainty for the symbol. Suppose that the sample space S has an alphabet size $\|S\| = N$. The entropy $H(X)$ attains the maximum value when the p.f. has a uniform distribution, i.e.:

$$P(x_i) = P(x_j) = \frac{1}{N} \quad \text{for all } i \text{ and } j \tag{3.176}$$

Equation (3.176) can be interpreted to mean that *uncertainty* reaches its maximum level when no outcome is more probable than any other. It can be proved that the entropy $H(X)$ is nonnegative and becomes zero only if the probability function is a deterministic one, i.e.,

$$H(X) \geq 0 \text{ with equality i.f.f. } P(x_i) = 1 \text{ for some } x_i \in S \tag{3.177}$$

There is another very interesting property for the entropy. If we replace the p.f. of generating symbol x_i in Eq. (3.175) with any other arbitrary p.f., the new value is no smaller than the original entropy. That is,

$$H(X) \le E\left[-\log Q(X)\right] = -\sum_{S} P(x_i)\log Q(x_i) \tag{3.178}$$

Equation (3.178) has a very important meaning. It shows that we are more uncertain about the data if we misestimate the distribution governing the data source. The equality for Eq. (3.178) occurs if and only if $P(x_i) = Q(x_i) \;\; 1 \le i \le N$. Equation (3.178), often referred to as *Jensen's inequality*, is the basis for the proof of EM algorithm in Chapter 4. Similarly, Jensen's ineqality can be extended to continuous pdf:

$$-\int f_x(x)\log f_x(x)dx \le -\int g_x(x)\log f_x(x)dx \tag{3.179}$$

with equality occurring if and only if $f_x(x) = g_x(x) \;\; \forall x$.

The proof of Eq. (3.178) follows from the fact $\log(x) \le x-1, \; \forall x$, so the following quantity must have an non-positive value.

$$\sum_{S} P(x_i)\log\frac{Q(x_i)}{P(x_i)} \le \sum_{S} P(x_i)\left[\log\frac{Q(x_i)}{P(x_i)} - 1\right] = 0 \tag{3.180}$$

Based on this property, the negation of the quantity in Eq. (3.180) can be used for the measurement of the distance of two probability distributions. Specifically, the *Kullback-Leibler* (KL) *distance* (*relative entropy*, *discrimination*, or *divergence*) is defined as:

$$KL(P \,||\, Q) = E\left[\log\frac{P(X)}{Q(X)}\right] = \sum_{S} P(x_i)\log\frac{P(x_i)}{Q(x_i)} \tag{3.181}$$

As discussed in Chapter 11, the branching factor of a grammar or language is an important measure of degree of difficulty of a particular task in spoken language systems. This relates to the size of the word list from which a speech recognizer or a natural language processor needs to disambiguate in a given context. According to the entropy definition above, this branching factor estimate (or average choices for an alphabet) is defined as follows:

$$PP(X) = 2^{H(X)} \tag{3.182}$$

$PP(X)$ is called the *perplexity* of source *X*, since it describes how confusing the grammar (or language) is. The value of perplexity is equivalent to the size of an imaginary equivalent list, whose words are equally probable. The bigger the perplexity, the higher branching factor. To find out the perplexity of English, Shannon devised an ingenious way [22] to estimate the entropy and perplexity of English words and letters. His method is similar to a guessing game where a human subject guesses sequentially the words of a text hidden from him, using the relative frequencies of her/his guesses as the estimates of the probability distribution underlying the source of the text. Shannon's perplexity estimate of English comes out to be about 2.39 for English letters and 130 for English words. Chapter 11 has a detailed description on the use of perplexity for language modeling.

3.4.2. Conditional Entropy

Now let us consider transmission of symbols through an information channel. Suppose the input alphabet is $X = (x_1, x_2, \ldots, x_s)$, the output alphabet is $Y = (y_1, y_2, \ldots, y_t)$, and the information channel is defined by the channel matrix $M_{ij} = P(y_j \mid x_i)$, where $P(y_j \mid x_i)$ is the conditional probability of receiving output symbol y_j when input symbol x_i is sent. Figure 3.17 shows an example of an information channel.

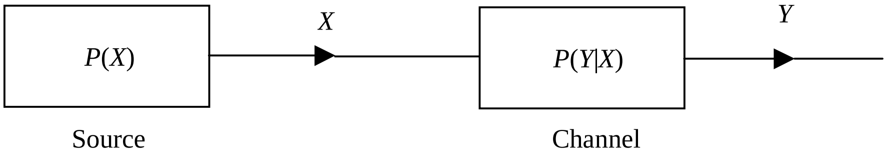

Figure 3.17 Example of information channel. The source is described by source p.f. *P*(*X*) and the channel is characterized by the conditional p.f. *P*(*Y*|*X*).

Before transmission, the average amount of information, or the uncertainty of the input alphabet *X*, is the prior entropy *H*(X).

$$H(X) = \sum_X P(X = x_i) \log \frac{1}{P(X = x_i)} \tag{3.183}$$

where $P(x_i)$ is the prior probability. After transmission, suppose y_j is received; then the average amount of information, or the uncertainty of the input alphabet A, is reduced to the following *posterior* entropy.

$$H(X \mid Y = y_j) = -\sum_X P(X = x_i \mid Y = y_j) \log P(X = x_i \mid Y = y_j) \tag{3.184}$$

where the $P(x_i \mid y_j)$ are the posterior probabilities. Averaging the posterior entropy $H(X \mid y_j)$ over all output symbols y_j leads to the following equation:

$$\begin{aligned} H(X \mid Y) &= \sum_Y P(Y = y_j) H(X \mid Y = y_j) \\ &= -\sum_Y P(Y = y_j) \sum_X P(X = x_i \mid Y = y_j) \log P(X = x_i \mid Y = y_j) \qquad (3.185) \\ &= -\sum_X \sum_Y P(X = x_i, Y = y_j) \log P(X = x_i \mid Y = y_j) \end{aligned}$$

This *conditional entropy*, defined in Eq. (3.185), is the average amount of information or the uncertainty of the input alphabet *X* given the outcome of the output event *Y*. Based on the definition of conditional entropy, we derive the following equation:

$$\begin{aligned} H(X,Y) &= -\sum_{X}\sum_{Y} P(X=x_i, Y=y_i)\log P(X=x_i, Y=y_i) \\ &= -\sum_{X}\sum_{Y} P(X=x_i, Y=y_i)\left\{\log P(X=x_i) + \log P(Y=y_i \mid X=x_i)\right\} \\ &= H(X) + H(Y \mid X) \end{aligned} \tag{3.186}$$

Equation (3.186) has an intuitive meaning – the uncertainty about two random variables equals the sum of uncertainty about the first variable and the conditional entropy for the second variable given the first variable is known. Equations (3.185) and (3.186) can be generalized to random vectors **X** and **Y** where each contains several random variables.

It can be proved that the chain rule [Eq. (3.16)] applies to entropy.

$$H(X_1,\cdots,X_n) = H(X_n \mid X_1,\cdots,X_{n-1}) + \cdots + H(X_2 \mid X_1) + H(X_1) \tag{3.187}$$

Finally, the following inequality can also be proved:

$$H(X \mid Y,Z) \le H(X \mid Y) \tag{3.188}$$

with equality i.f.f. X and Z being independent when conditioned on Y. Equation (3.188) basically confirms the intuitive belief that uncertainty decreases when more information is known.

3.4.3. The Source Coding Theorem

Information theory is the foundation for data compressing. In this section we describe *Shannon's source coding theorem*, also known as the *first coding theorem*. In source coding, we are interested in *lossless* compression, which means the compressed information (or symbols) can be recovered (decoded) perfectly. The entropy serves as the upper bound for a source lossless compression.

Consider an information source with alphabet $S = \{0,1,\ldots,N-1\}$. The goal of data compression is to *encode* the output symbols into a string of binary symbols. An interesting question arises: *What is the minimum number of bits required, on the average, to encode the output symbols of the information source*?

Let's assume we have a source that can emit four symbols {0,1,2,3} with equal probability $P(0) = P(1) = P(2) = P(3) = 1/4$. Its entropy is 2 bits as illustrated in Eq. (3.189):

$$H(S) = \sum_{i=0}^{3} P(i)\log_2 \frac{1}{P(i)} = 2 \tag{3.189}$$

It is obvious that 2 bits per symbol is good enough to encode this source. A possible binary code for this source is {00, 01, 10, 11}. It could happen, though some symbols are more likely than others, for example, $P(0)=1/2,\ P(1)=1/4,\ P(2)=1/8,\ P(3)=1/8$. In this case the entropy is only 1.75 bits. One obvious idea is to use fewer bits for lower values that are frequently used and more bits for larger values that are rarely used. To represent this source we can use a variable-length code {0,10,110,111}, where no codeword is a prefix for the rest and thus a string of 0's and 1's can be uniquely broken into those symbols. The encoding scheme with such a property is called *uniquely decipherable* (or instantaneous) coding, because as soon as the decoder observes a sequence of codes, it can decisively determine the sequence of the original symbols. If we let $r(x)$ be the number of bits (length) used to encode symbol *x*, the average rate *R* of bits per symbol used for encoding the information source is:

$$R = \sum_x r(x)P(x) \tag{3.190}$$

In our case, *R* is 1.75 bits as shown in Eq. (3.191):

$$R = 0.5\times 1 + 0.25\times 2 + 0.125\times 3 + 0.125\times 3 = 1.75 \tag{3.191}$$

Such variable-length coding strategy is called *Huffman coding*. Huffman coding belongs to *entropy coding* because it matches the entropy of the source. In general, *Shannon's source coding* theorem says that a source cannot be coded with fewer bits than its entropy. We will skip the proof here. Interested readers can refer to [3, 15, 17] for the detailed proof. This theorem is consistent with our intuition because the entropy measure is exactly the information content of the information measured in bits. If the entropy increases, then uncertainty increases, resulting in a large amount of information. Therefore, it takes more bits to encode the symbols. In the case above, we are able to match this rate, but, in general, this is impossible, though we can get arbitrarily close to it. The Huffman code for this source offers a compression rate of 12.5% relative to the code designed for the uniform distribution.

Shannon's source coding theorem establishes not only the lower bound for lossless compression but also the upper bound. Let $\lceil x \rceil$ denote the smallest integer that greater or equal to *x*. As in the similar procedure above, we can make the code length assigned to source output x equal to

$$l(x) = \lceil -\log P(x) \rceil \tag{3.192}$$

The average length *L* satisfies the following inequality:

$$L = \sum_x l(x)P(x) < \sum_x \left[1 - \log P(x)\right] P(x) = 1 + H(X) \tag{3.193}$$

Equation (3.193) means that the average rate *R* only exceeds the value of entropy by less than one bit.

L can be made arbitrary close to the entropy by block coding. Instead of encoding single output symbols of the information source, one can encode each block of length *n*. Let's assume the source is memoryless, so $X_1, X_2, \ldots, X_n$ are independent. According to Eq. (3.193), the average rate *R* for this block code satisfies:

$$L < 1 + H(X_1, X_2, \ldots, X_n) = 1 + nH(X) \tag{3.194}$$

This makes the average number of bits per output symbol, L/n, satisfy

$$\lim_{n\to\infty}\frac{1}{n}L \le H(X) \tag{3.195}$$

In general, Huffman coding arranges the symbols in order of decreasing probability, assigns the bit 0 to the symbol of highest probability and the bit 1 to what is left, and proceeds the same way for the second highest probability value (which now has a code 10) and iterate. This results in 2.25 bits for the uniform distribution case, which is higher than the 2 bits we obtain with equal-length codes.

Lempel-Ziv coding is a coding strategy that uses correlation to encode *strings* of symbols that occur frequently. Although it can be proved to converge to the entropy, its convergence rate is much slower [27]. Unlike Huffman coding, Lempel-Ziv coding is independent of the distribution of the source; i.e., it needs not be aware of the distribution of the source before encoding. This type of coding scheme is often referred to as *universal* encoding scheme.

3.4.4. Mutual Information and Channel Coding

Let's review the information channel illustrated in Figure 3.17. An intuitively plausible measure of the average amount of information provided by the random event *Y* about the random event *X* is the average difference between the number of bits it takes to specify the outcome of *X* when the outcome of *Y* is not known and the outcome of Y is known. *Mutual information* is defined as the difference in the entropy of *X* and the conditional entropy of *X* given *Y*:

$$\begin{aligned}
I(X;Y) &= H(X) - H(X\,|\,Y) \\
&= \sum_X P(x_i)\log\frac{1}{P(x_i)} - \sum_X\sum_Y P(x_i,y_j)\log\frac{1}{P(x_i\,|\,y_j)} \\
&= \sum_X\sum_Y P(x_i,y_j)\log\frac{P(x_i\,|\,y_j)}{P(x_i)} = \sum_X\sum_Y P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(x_i)P(y_i)} \\
&= E\left[\log\frac{P(X,Y)}{P(X)P(Y)}\right]
\end{aligned} \tag{3.196}$$

$I(X;Y)$ is referred to as the mutual information between *X* and *Y*. $I(X;Y)$ is symmetrical; i.e., $I(X;Y) = I(Y;X)$. The quantity $P(x,y)/P(x)P(y)$ is often referred to as the *mutual information between symbol x and y*. $I(X;Y)$ is bounded:

$$0 \le I(X;Y) \le \min\left[H(X), H(Y)\right] \tag{3.197}$$

$I(X;Y)$ reaches the minimum value (zero) when the random variables *X* and *Y* are independent.

Mutual information represents the information obtained (or the reduction in uncertainty) through a channel by observing an output symbol. If the information channel is noiseless, the input symbol can be determined definitely by observing an output symbol. In this case, the conditional entropy $H(X|Y)$ equals zero and it is called a *noiseless channel*. We obtain the maximum mutual information $I(X; Y) = H(X)$. However, the information channel is generally *noisy* so that the conditional entropy $H(X|Y)$ is not zero. Therefore, maximizing the mutual information is equivalent to obtaining a low-noise information channel, which offers a closer relationship between input and output symbols.

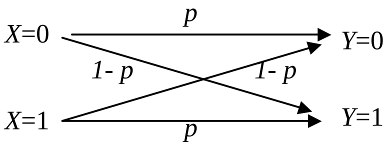

Figure 3.18 A binary channel with two symbols.

Let's assume that we have a binary channel, a channel with a binary input and output. Associated with each output are a probability p that the output is correct, and a probability $(1-p)$ that it is not, so that the channel is *symmetric*.

If we observe a symbol Y = 1 at the output, we don't know for sure what symbol X was transmitted, though we know $P(X=1|Y=1)=p$ and $P(X=0|Y=1)=(1-p)$, so that we can measure our uncertainty about X by its conditional entropy:

$$H(X|Y=1) = -p\log p - (1-p)\log(1-p) \tag{3.198}$$

If we assume that our source X has a uniform distribution, $H(X|Y) = H(X|Y=1)$ as shown in Eq. (3.198) and $H(X)$ = 1. The mutual information between X and Y is given by

$$I(X,Y) = H(X) - H(X|Y) = 1 + p\log p + (1-p)\log(1-p) \tag{3.199}$$

It measures the information that Y carries by about X. The channel capacity C is the maximum of the mutual information over all distributions of X. That is,

$$C = \max_{P(x)} I(X;Y) \tag{3.200}$$

The channel capacity C can be attained by varying the distribution of the information source until the mutual information is maximized for the channel. The channel capacity C can be regarded as a channel that can transmit at most C bits of information per unit of time. *Shannon's channel coding* theorem says that for a given channel there exists a code that permits error-free transmission across the channel, provided that $R \le C$, where R is the rate of the communication system, which is defined as the number of bits per unit of time being transmitted by the communication system. Shannon's channel coding theorem states the fact that *arbitrarily reliable communication is possible at any rate below channel capability*.

Figure 3.19 illustrates a transmission channel with the source decoder and destination decoder. The source encoder will encode the source symbol sequence $\mathbf{x} = x_1, x_2, \ldots, x_n$ into channel input sequence $y_1, y_2, \ldots, y_k$. The destination decoder takes the output sequence $z_1, z_2, \ldots, z_k$ from the channel and converts it into the estimates of the source output

$\overline{\mathbf{x}} = \overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n$. The goal of this transmission is to make the probability of correct decoding $P(\overline{\mathbf{x}} = \mathbf{x})$ asymptotically close to 1 while keeping the compression ratio $\Re = n/k$ as large as possible. *Shannon's source-channel coding* theorem (also referred to as *Shannon's second coding* theorem) says that it is possible to find an encoder-decoder pair of rate $\Re$ for a noisy information channel, provided that $\Re \times H(X) \le C$.

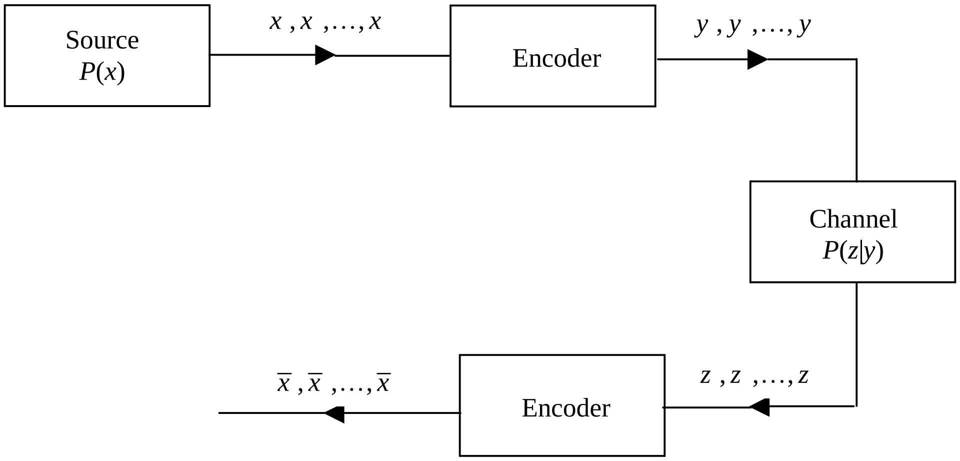

Figure 3.19 Transmission of information through a noisy channel [15].

Because of channel errors, speech coders need to provide error correction codes that will decrease the bit rate allocated to the speech. In practice, there is a tradeoff between the bit rate used for source coding and the bit rate for channel coding. In Chapter 7 we will describe speech coding in great detail.

3.5. HISTORICAL PERSPECTIVE AND FURTHER READING

The idea of uncertainty and probability can be traced all the way back to about 3500 B.C., when games of chance played with bone objects were developed in Egypt. Cubical dice with markings virtually identical to modern dice have been found in Egyptian tombs dating in around 2000 B.C. Gambling with dice played an important part in the early development of probability theory. Modern mathematical theory of probability is believed to have been started by the French mathematicians Blaise Pascal (1623-1662) and Pierre Fermat (1601-1665) when they worked on certain gambling problems involving dice. English mathematician Thomas Bayes (1702-1761) was first to use probability inductively and established a mathematical basis for probability inference, leading to what is now known as Bayes' theorem. The theory of probability has developed steadily since then and has been widely applied in diverse fields of study. There are many good textbooks on probability theory. The book by DeGroot [6] is an excellent textbook for both probability and statistics which covers all the necessary elements for engineering majors. The authors also recommend [14], [19], or [24] for interested readers.

Estimation theory is a basic subject in statistics covered in textbooks. The books by DeGroot [6], Wilks [26] and Hoel [13] offer excellent discussions of estimation theory. They all include comprehensive treatments for maximum likelihood estimation and Bayesian estimation. Maximum likelihood estimation was introduced by in 1912 R. A. Fisher (1890-1962) and has been applied to various domains. It is arguably the most popular parameter estimation method due to its intuitive appeal and excellent performance with large training samples. The EM algorithm in Chapter 4 and the estimation of hidden Markov models in Chapter 8 are based on the principle of MLE. The use of prior distribution in Bayesian estimation is very controversial in statistics. Some statisticians adhere to the Bayesian philosophy of statistics by taking the Bayesian estimation' view of the parameter Φ having a probability distribution. Others, however, believe that in many problems Φ is not a random variable but rather a fixed number whose value is unknown. Those statisticians believe that a prior distribution can be assigned to a parameter Φ only when there is extensive prior knowledge of the past; thus the non-informative priors are completely ruled out. Both groups of statisticians agree that whenever a meaningful prior distribution can be obtained, the theory of Bayesian estimation is applicable and useful. The books by DeGroot [6] and Poor[20] are excellent for learning the basics of Bayesian and MAP estimations. Bayesian and MAP adaptation are particularly powerful when the training samples are sparse. Therefore, they are often used for adaptation where the knowledge of prior distribution can help to adapt the model to a new but limited training set. The speaker adaptation work done by Brown et al. [2] first applied Bayesian estimation to speech recognition and [9] is another good paper on using MAP for hidden Markov models. References [4], [16] and [14] have extensive studies of different conjugate prior distributions for various standard distributions. Finally, [1] has an extensive reference for Bayesian estimation.

Significance testing is an essential tool for statisticians to interpret all the statistical experiments. Neyman and Person provided some of the most important pioneering work in hypotheses testing [18]. There are many different testing methods presented in most statistics book. The χ^2 test, invented in 1900 by Karl Pearson, is arguably the most widely used testing method. Again, the textbook by DeGroot [6] is an excellent source for the basics of testing and various testing methods. The authors recommend [7] as an interesting book that uses many real-world examples to explain statistical theories and methods, particularly the significance testing.

Information theory first appeared in Claude Shannon's historical paper: *A Mathematical Theory of Communication* [21]. In it, Shannon, analyzed communication as the transmission of a message from a source through a channel to a receiver. In order to solve the problem he created a new branch of applied mathematics - *information and coding theory*. IEEE published a collection of Shannon's papers [23] containing all of his published works, as well as many that have never been published. Those published include his classic papers on information theory and switching theory. Among the unpublished works are his once-secret wartime reports, his Ph.D. thesis on population genetics, unpublished Bell Labs memoranda, and a paper on the theory of juggling. The textbook by McEliece [17] is excellent for learning all theoretical aspects of information and coding theory. However, it might be out of print now. Instead, the books by Hamming [12] and Cover [3] are two current great references for information and coding theory. Finally, F. Jelinek's *Statistical Methods for Speech Recognition* [15] approaches the speech recognition problem from an information-theoretic aspect. It is a useful book for people interested in both topics.

REFERENCES

[1] Bernardo, J.M. and A.F.M. Smith, *Bayesian Theory*, 1996, New York, John Wiley.

[2] Brown, P., C.-H. Lee, and J. Spohrer, "Bayesian Adaptation in Speech Recognition," *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1983, Boston, MA pp. 761-764.

[3] Cover, T.M. and J.A. Thomas, *Elements of Information Theory*, 1991, New York, John Wiley and Sons.

[4] DeGroot, M.H., *Optimal Statistical Decisions*, 1970, New York, NY, McGraw-Hill.

[5] DeGroot, M.H., *Probability and Statistics*, Addison-Wesley Series in Behavioral Science: Quantitive Methods, eds. F. Mosteller, 1975, Reading, MA, Addison-Wesley Publishing Company.

[6] DeGroot, M.H., *Probability and Statistics*, 2nd ed, Addison-Wesley Series in Behavioral Science: Quantitive Methods, eds. F. Mosteller, 1986, Reading, MA, Addison-Wesley Publishing Company.

[7] Freedman, D., *et al.*, *Statistics*, 2nd ed, 1991, New York, W. W. Norton & Company, Inc.

[8] Gales, M.J., *Model Based Techniques for Noise Robust Speech Recognition*, PhD Thesis in *Engineering Department* 1995, Cambridge University, .

[9] Gauvain, J.L. and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, 1994, **2**(2), pp. 291-298.

[10] Gillett, G.E., *Introduction to Operations Research: A Computer-Oriented Algorithmic Approach*, McGraw-Hill Series in Industrial Engineering and Management Science, eds. J. Riggs, 1976, New York, McGraw-Hill.

[11] Gillick, L. and S.J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1989, Glasgow, Scotland, UK, IEEE pp. 532-535.

[12] Hamming, R.W., *Coding and Information Theory*, 1986, Englewood Cliffs NJ, Prentice-Hall.

[13] Hoel, P.G., *Introduction to Mathesmatical Statistics*, 5th edition ed, 1984, John Wiley & Sons.

[14] Jeffreys, H., *Theory of Probability*, 1961, Oxford University Press.

[15] Jelinek, F., *Statistical Methods for Speech Recognition*, Language, Speech, and Communication, 1998, Cambridge, MA, MIT Press.

[16] Lindley, D.V., "The Use of Prior Probability Distributions in Statistical Inference and Decision," *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, Berkeley, CA, Univ. of California Press.

[17] McEliece, R., *The Theory of Information and Coding*, Encyclopedia of Mathematics and Its Applications, eds. R. Gian-Carlo. Vol. 3, 1977, Reading, Addison-Wesley Publishing Company.

[18] Neyman, J. and E.S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Trans. of Royal Society*, 1928, **231**, pp. 289-337.

[19] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, 3rd ed, 1991, New York, McGraw-Hill.

[20] Poor, H.V., *An Introduction to Signal Detection and Estimation*, Springer tests in Electrical Engineering, eds. J.B. Thomas, 1988, New York, Springer-Verlag.

[21] Shannon, C., "A Mathematical Theory of Communication System," *Bell System Technical Journal*, 1948, **27**, pp. 379-423, 623-526.

[22] Shannon, C.E., "Prediction and Entropy of Printed English," *Bell System Technical Journal*, 1951, pp. 50-62.

[23] Shannon, C.E., *Claude Elwood Shannon : Collected Papers*, 1993, IEEE.

[24] Viniotis, Y., *Probability and Random Processes for Electrical Engineering*, Outline Series in Electronics & Electrical Engineering, eds. Schaum, 1998, New York, WCB McGraw-Hill.

[25] Wald, A., "Note of Consistency of Maximum Likelihood Estimate," *Ann. Mathematical Statistics*, 1949(20), pp. 595-601.

[26] Wilks, S.S., *Mathematical Statistics*, 1962, New York, John Wiley and Sons.

[27] Ziv, J. and A. Lempel, "A Universal Algorithm for Sequential Data Compression," *IEEE Trans. on Information Theory*, 1997, **IT-23**, pp. 337-343.